Localization in e-Governance Perspective

WELCOME
सुस्वागतम्
நல்வரவு
ସୁସ୍ୱାଗତଂ
సుస్వాగతం
ಸುಸ್ವಾಗತಂ
സുസ്വാഗതം
সুস্বাগতম
ਸੁਆਗਤਮ
સુસ્વાગતમ્
خوش آمدید

Mahesh D. Kulkarni
Associate Director & HoD, GIST
C-DAC, Pune
mdk@cdac.in

Venue :
Hotel Novatel

27th April 2013

Outline

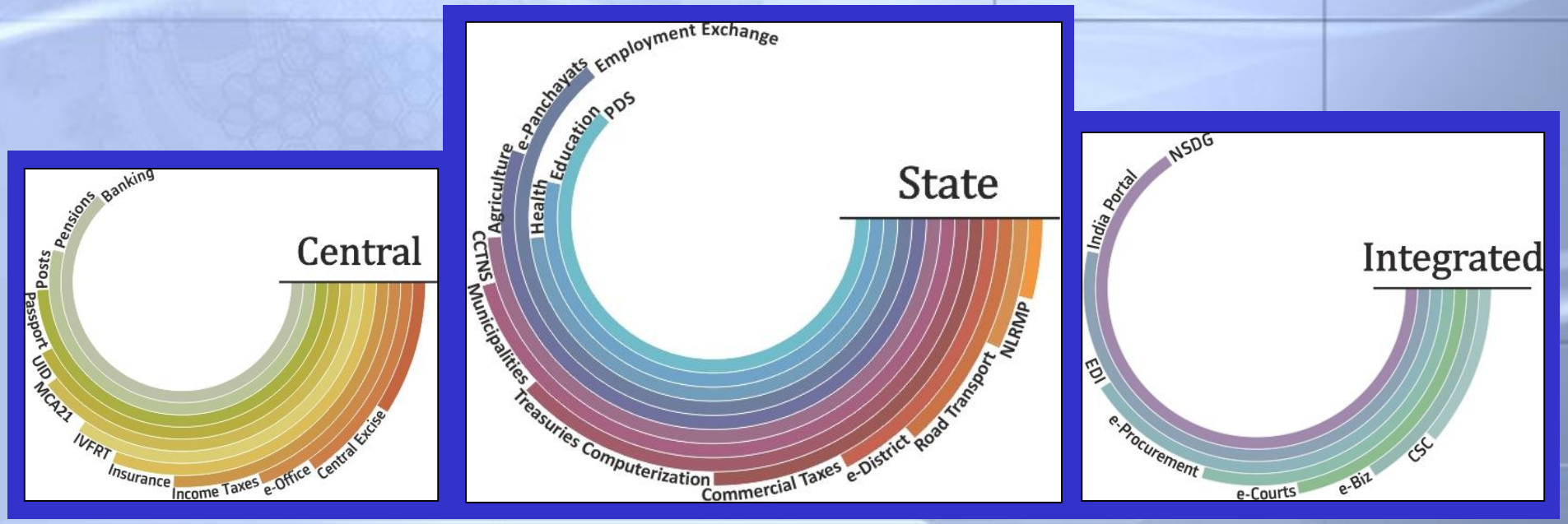
- Importance of localization
- Internationalization / localization
- The game changer – Standards
- Indian standards – GIGW, e-governance localization best practices, (RFP)
- Various standards in context of Indian languages
- Surface level localization – what it takes
- Deeper level localization – what it takes (example of databases)

CDAC offering :

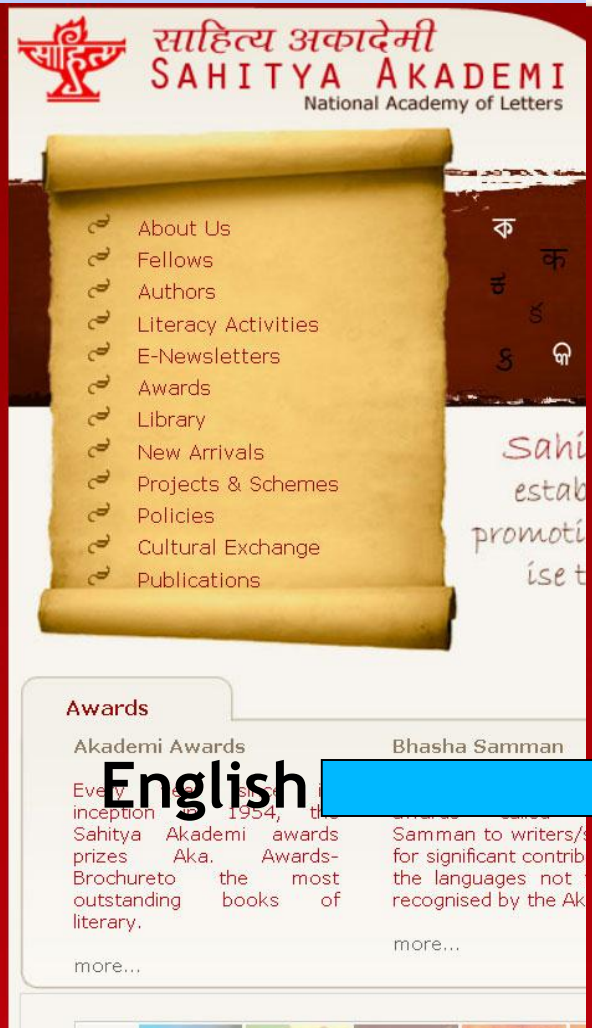
- Localization Framework
- Leveraging on prior knowledge – collaborative model

Mission Mode Projects

NeGP comprises of 31 Mission Mode Projects (MMPs) encompassing 11 Central MMPs, 13 State MMPs and 7 Integrated MMPs spanning multiple Ministries/ Departments.



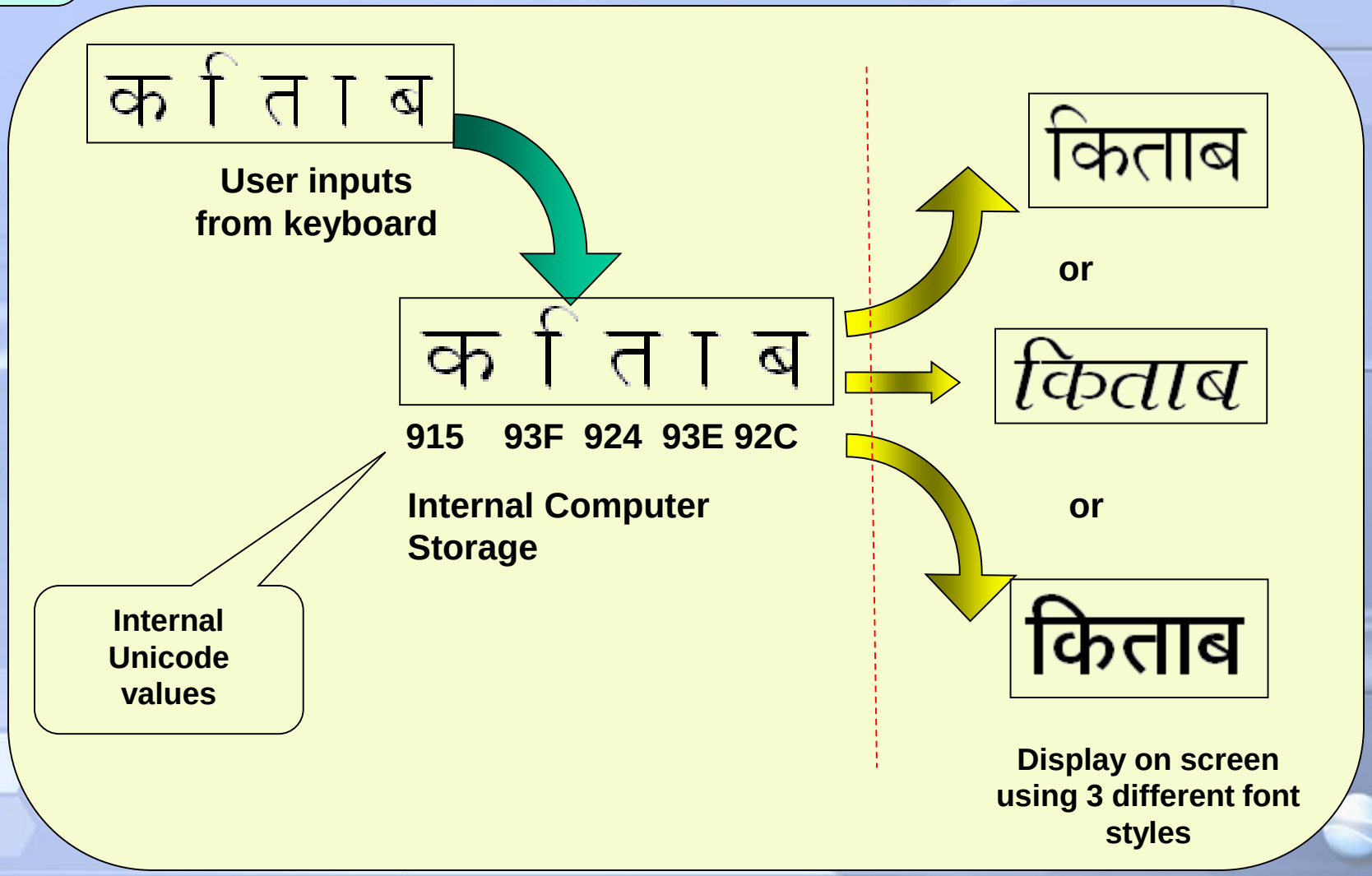
Hence Indian language interface becomes crucial for easy access to the information



English → Hindi

कि

Akshara formation



Language nuances

- One script – Many language (Devanagari script – Hindi, Marathi, Konkani, Nepali, Sindhi, Kashmiri, Dogri, Boro, Santali)
- Many language – one script (Santali language – Devanagari, OlChiki

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

अनून : 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)

What needs to be done

Follow internationalization best practices which will help

- To build globally acceptable software
- Isolate translatable and non-translatable resources
- Universal encoding
- Interoperability
- Easy localisation

At MMP Level

- Include best practices in RFP
- Ensure adherence of these standards at every stage of MMP development
- Train stakeholders / developers in “localisation”

Best practices to be included RFP

- GIGW – Government of Indian guidelines for Web
- e-Governance localization guidelines
- W3C standards
- WCAG
- Translation style guides

Internationalization & Localization

Internationalization

- Internationalization is the process of designing a software application so that it can be adapted to various languages and regions by reducing engineering efforts

Localization :

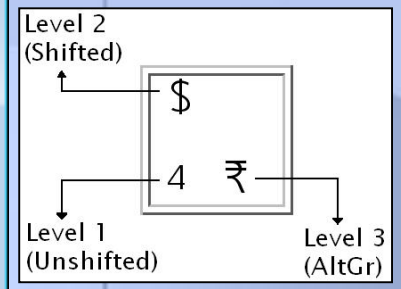
- Adaptation of a product, application or document content to meet the language, cultural and other requirements of a specific region (a "locale").
- Apart from translation of the UI and documentation, it covers
  - Numbers, date, time formats, currency, keyboard, Collation, Symbols, icons, colors and many more.

Standards

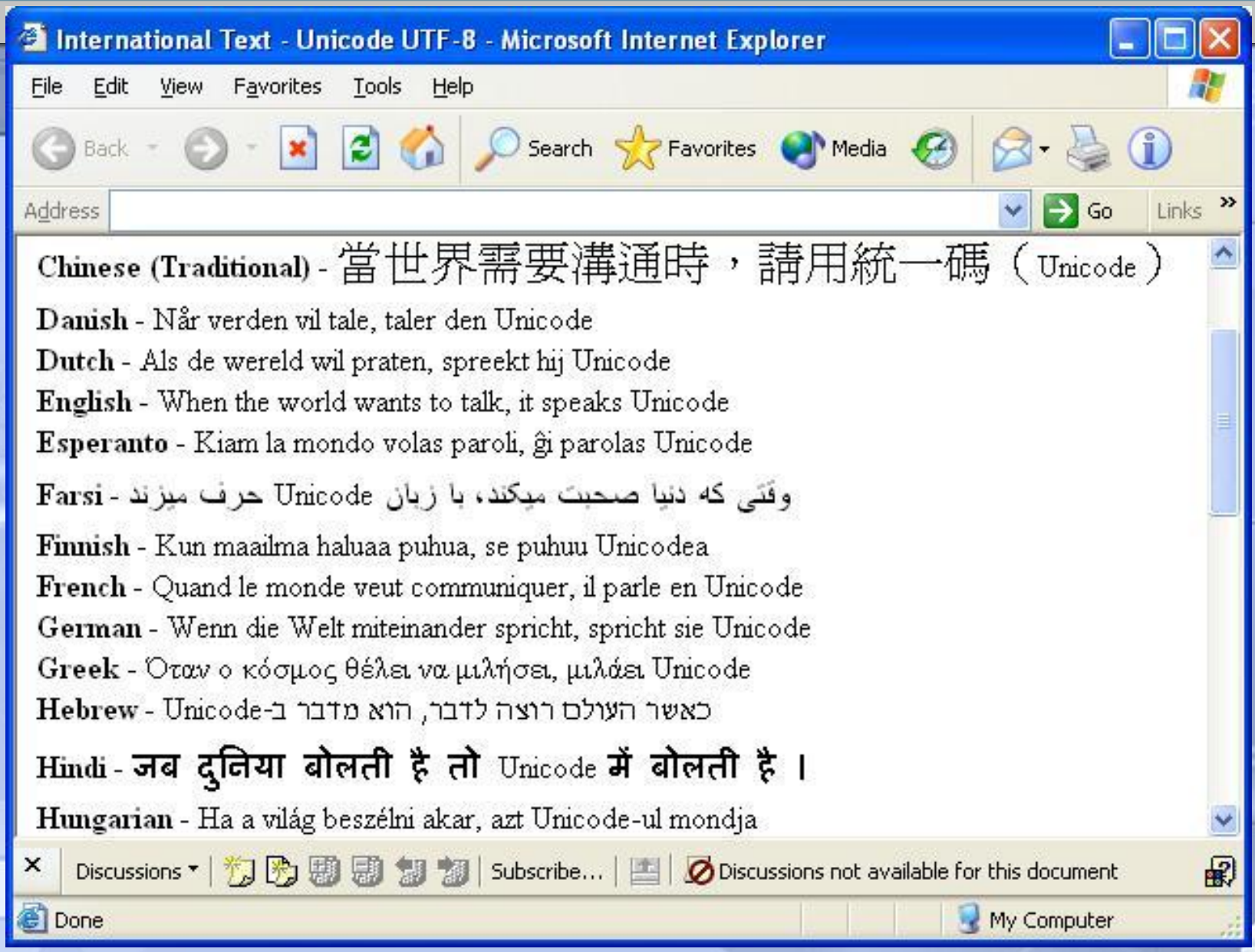
- Inputting standards – Enhanced INSCRIPT Keyboard. (Onscreen Keyboard for apps)
  - Stylus based devices (tablets / mobiles) – either INSCRIPT keyboard and or 12 keys Virtual Keyboard
  - Non-stylus based devices (tablets / mobiles) – physical INSCRIPT or 12 keys keyboard.
- Storage (Universal encoding – UTF8 / Unicode)
- Display - Unicode compliant "Open Font Format" – on lines of SakalBharati.
- FUEL – Frequently Used Entry List
- Term banks, Domain specific terms
- CLDR – Common Locale Data Repository
- Roman phonetic equivalents.
- W3C / WCAG *(Web Content Accessibility Guidelines)*
- Desirable to have Content Management System at backend for all the MMPs – for ease of localization.

Enhanced INSCRIPT Keyboard

Rupee Symbol



Unique representation globally - Unicode



Example of a multilingual Web page encoded in UTF-8 (Source: Microsoft).

Unicode compliant “Open Font Format” - SakalBharati

Sakal Bharat **Assamese** Normal

কঁহুৱা বন মোৰ অশান্ত মন,

Sakal Bharat **Devanagari** Normal

भारतीय भाषाएं और सामाजीक परिवर्तन।

Sakal Bharat **Gujarati** Normal

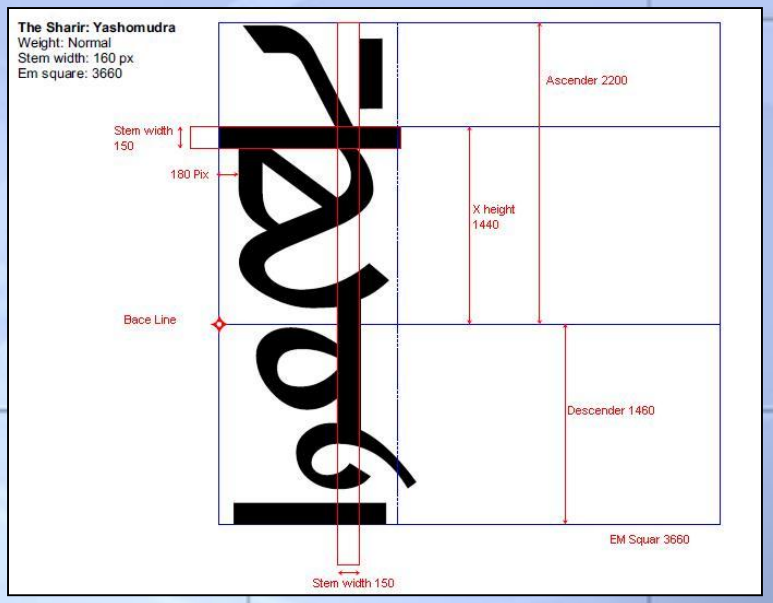
ભારતીય ભાષાઓ અને સામાજીક પરિવર્તન।

Sakal Bharat **Oriya** Normal

ଇଂଗିତ ପ୍ରଶ୍ନ ଦୟାକରି ମନୋନୀତ କରନ୍ତୁ

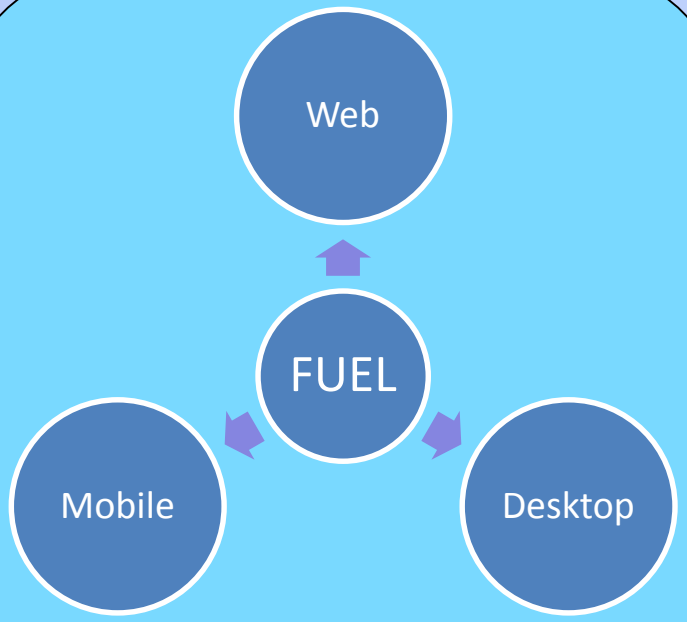
Sakal Bharat **Punjabi** Normal

ਪੰਜਾਬ ਪੰਜ ਦਰਿਆਵਾਂ ਦੀ ਧਰਤੀ ਹੈ।



کنپیوٹر کی دوڑ میں ہم کہاں ہیں؟

Frequently Used Entry List



FUEL is an open source initiative to standardize terms for open source software programs. It aims at resolving the problem of term inconsistency and lack of standardization in Computer software translation, across various platforms.

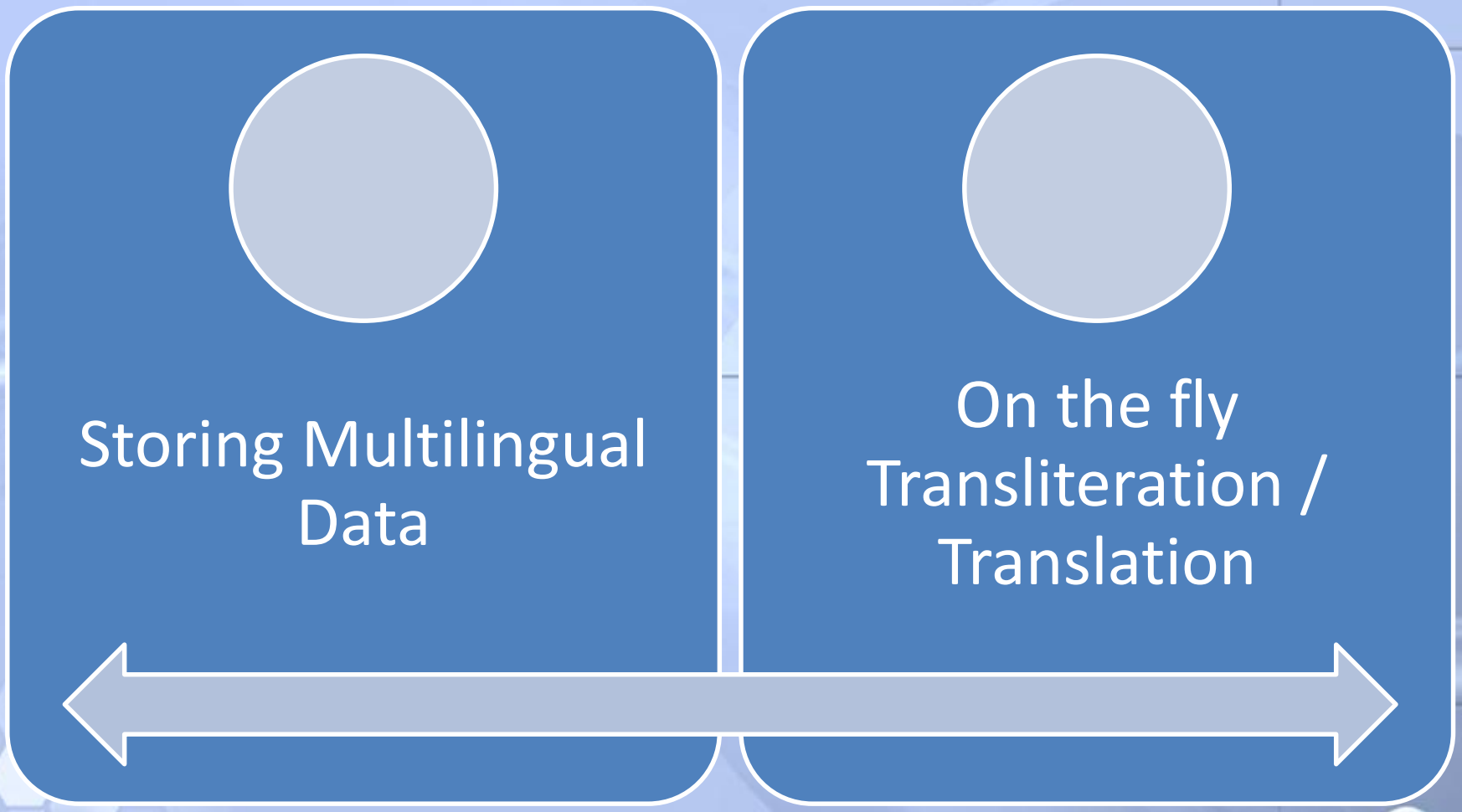
Common Locale Data Repository

Common Locale Data Repository - Example

```
<monthWidth type="wide">
<month type="1">जनवरी </month>
<month type="2"> फरवरी </month>
<month type="3">मार्च</month>
<month type="4">एप्रैल </month>
<month type="5">मेई </month>
<month type="6">जून </month>
<month type="7">जूलै </month>
<month type="8">अगस्त </month>
<month type="9">सितंबर </month>
<month type="10">अक्तूबर
</month>
<month type="11">नवंबर </month>
<month type="12">दिसंबर
</month>
```

For further details please refer:
http://*cldr.unicode.org/*

Languages & Databases



Searching in Databases

Alternate Spellings

- In Indian Languages several words have multiple correct spellings and alternate representation forms

Indian Language Numerals

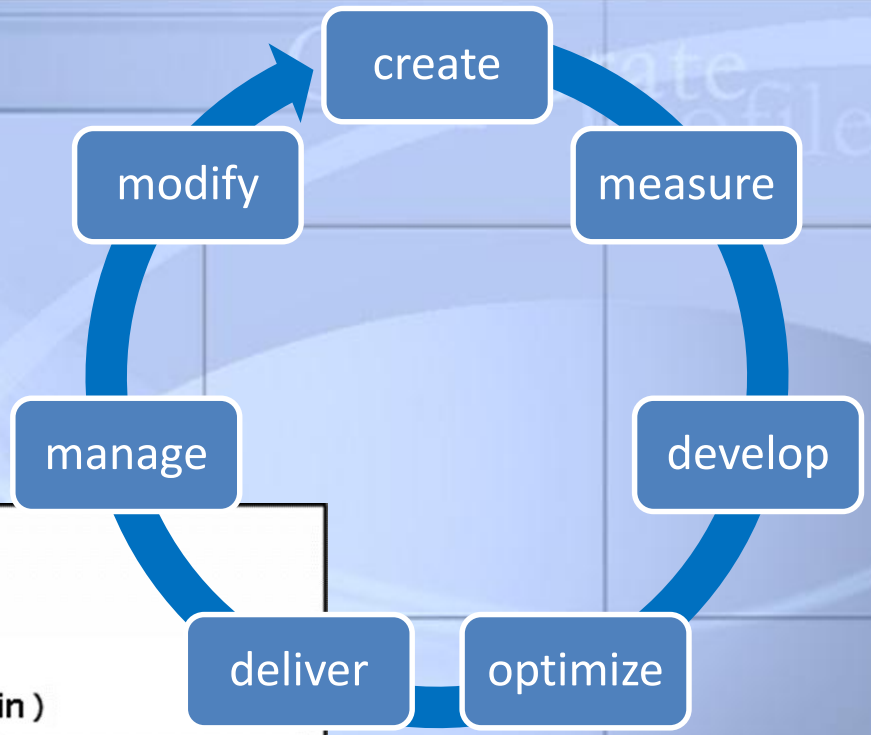
- IL numerals are not mapped to English numerals.

Sort order

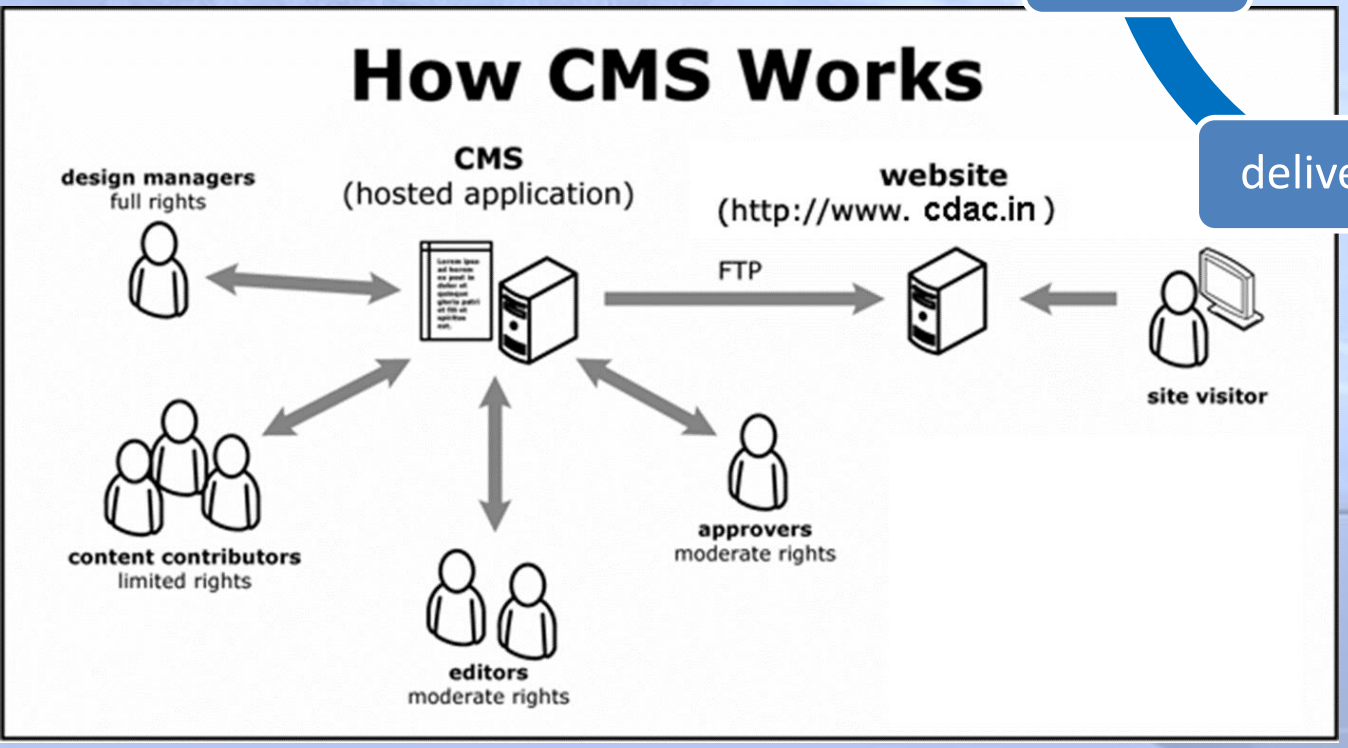
SORT ORDER : Sort order is affected by language.

As your can see that Hindi and English languages have different sorting order

Content Management System



Helps in easy and faster localisation



Search engine - Relevancy

- 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.

How CDAC can help?

Need of LPMF on Cloud platform

- Hurricane task to enable all the MMPs with region specific Indian languages if the architecture is not internationalized.
- Difficult to have seamless access to information without standards and process flow
- For SIM (SIMultaneous) release in multiple languages, application developers, translators, evaluators need to work in tandem.
- Volume, last minute localization
- Non-availability of skilled manpower, especially in the field of language translation

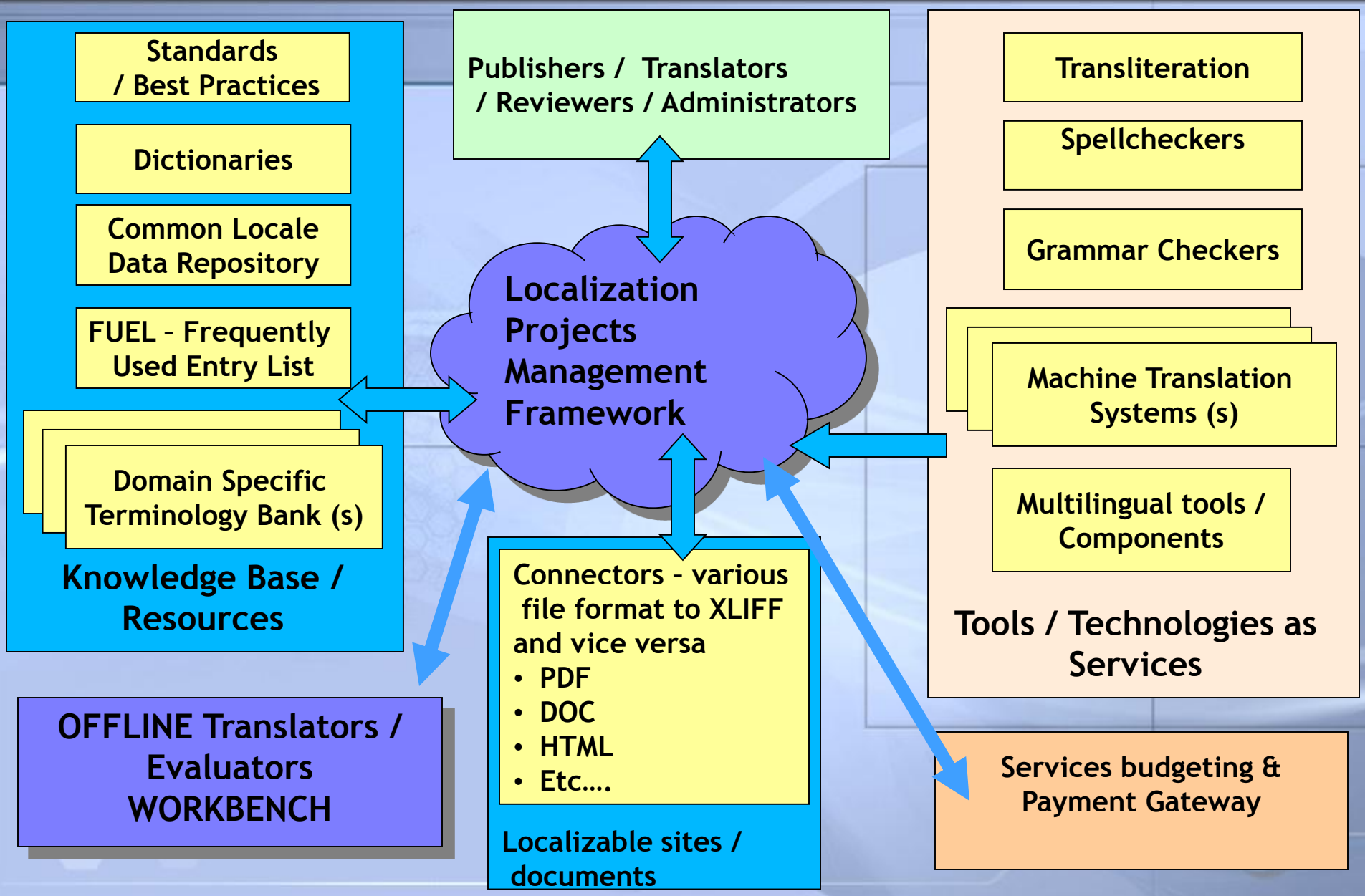
Workflow software can greatly streamline the localization process by managing the sequence of review/edit tasks, providing the status of tasks and processes, and notifying participants of changes in state, new work, or other information

What is needed ?

- **Unified standards – Unicode, XLIFF, TMX, SBX, CLDR, etc**
- **Localization Projects Management Framework offered as CLOUD service.**
- **Linguistic tools and resources**
  - Term Banks, Dictionaries, CLDR, FUEL
  - Spellchecker, grammar Checker, Transliteration tools
  - Search engines, Machine Translation System

  Entire Work Flow Management
  - Publisher, Translator, evaluator, administrators modules managements
- **Services offering and budgeting**
  - Document, books, applications localization services
- **Community participation**
  - Crowd sourcing, knowledge sharing, chat, RSS feeds, etc
- **Payment Gateways**
  - For payments to translators, evaluators, etc
- **Facility to plug-in Third party tools / technologies.**

Localization in Cloud



Benefits

- Workflow optimization
- Knowledge Repository
- Standards usage - XLIFF
- Machine Translation System (Consortia MT engine)
- Synchronization of various components
- For surface localization no need of source code (on the fly)
- Bringing translators, evaluators, Project Managers, publishers, tools provider on a single platform.
- Workflow saves money, time and infrastructure.
- Believes on leveraging prior / existing knowledge

Transliteration / Translation Toolbar

iPlugin | Highlight | Scriptkey | Keyboard | Floating Keyboard | Help | Convert

EnglishToHindi
HindiToUrdu
UrduToHindi

BBC

Home | News | Sport | Radio | TV | Weather | Languages

BBC HINDI.com

ڈھونڈیں

ہمارے کارپکرم سنیں ریڈیو پلیئر چلائیں

مدد چاہئے اسے بوکمارک بنائیں

پہلا پنّا
بھارت اور پڑوس
کھیل کی دنیا
منورنجن ایکسپریس
آپکی رائے
کچھ اور جانئے
Learning English

ریڈیو
ہمارے کاریہ کرم
برسارن سمے

سنیوکتی
ہم کون ہیں
ہمارا پتا

گرینتا

ٹک کے سماچار GMT منگلوار, 07 اگست, 2007 کو 05:48

باڑھ پیڑتوں تک راحت پہنچانے میں مشکلیں

بھارت سمیت دکشن ایشیا کے کئی دیشوں میں باڑھ سے پربھاوت لوگوں تک راحت سامگری پہنچانے میں راحتکرمیوں کو نقلوں کا سامنا کرنا پڑ رہا ہے۔

+ ناؤ ڈرگھٹنا میں 15 سے ادھیک موتیں

+ 'قسمت میں یہی لکھا ہے'...

ہندی فونٹ ڈاؤنلوڈ

यदि आप इस पेज को हिंदी में नहीं देख पा रहे हैं तो कृपया यहां क्लिक करें

मदद के लिए क्लिक करें

- ممبئی ٹرین دھماکے میں 13 کے خلاف آروپ
- "خطروں سے نپٹنے کے لئے تیار بھارت"
- پانچ اور سنّی منتریوں نے شپتھ لی

'تالیبان کو سماپت کرنے پر سہمتی'

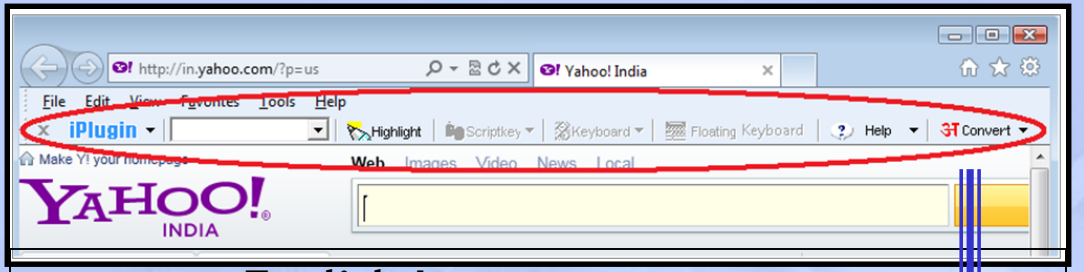
ٹوینٹی-20 ورلڈ کپ کے لئے بھارتیہ ٹیم

سلاخوں کے پیچھے سنجے

ہروشیما بم ہملے کی 62ویں برسی

دیکھئے

Surface localizer plug-in for various browsers

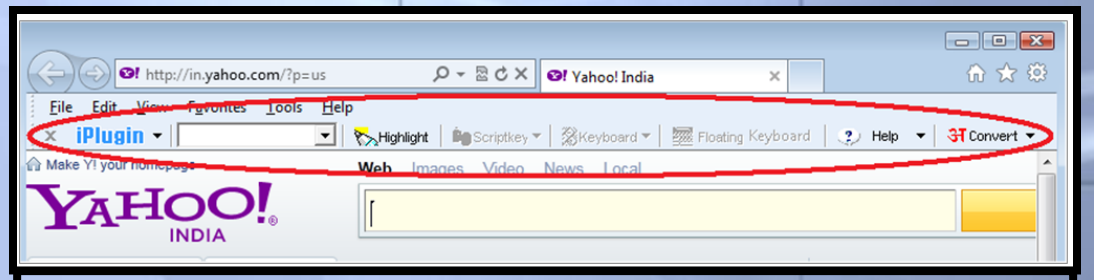


English Language contents

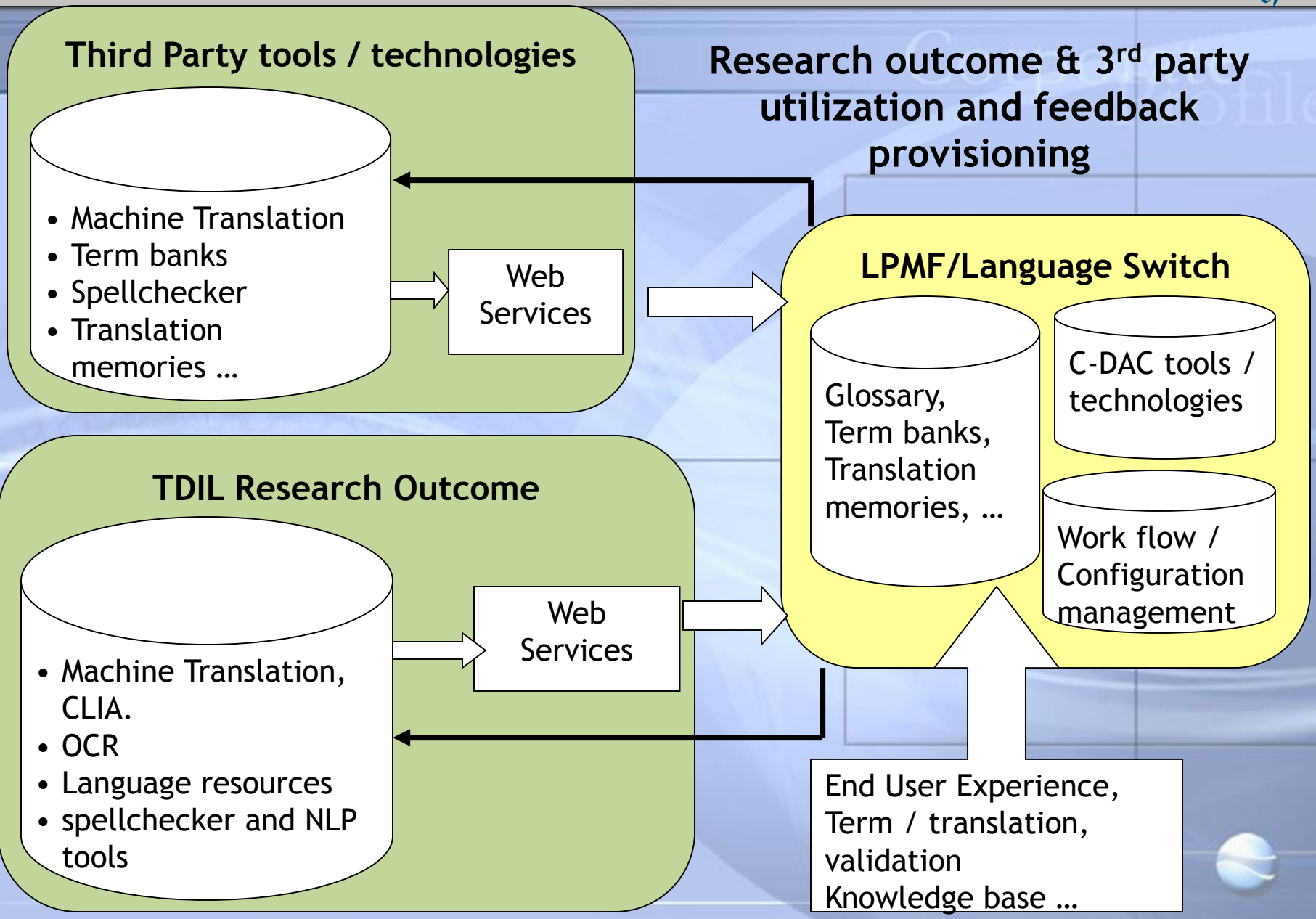
The plug-in invokes various web services from within, such as MAT, Spell check, Dictionaries, URL specific Term banks, Translation memories

Web Services

- Machine Translation, Term banks, Translation memories, …
- C-DAC tools / technologies
- Work flow / Configuration management

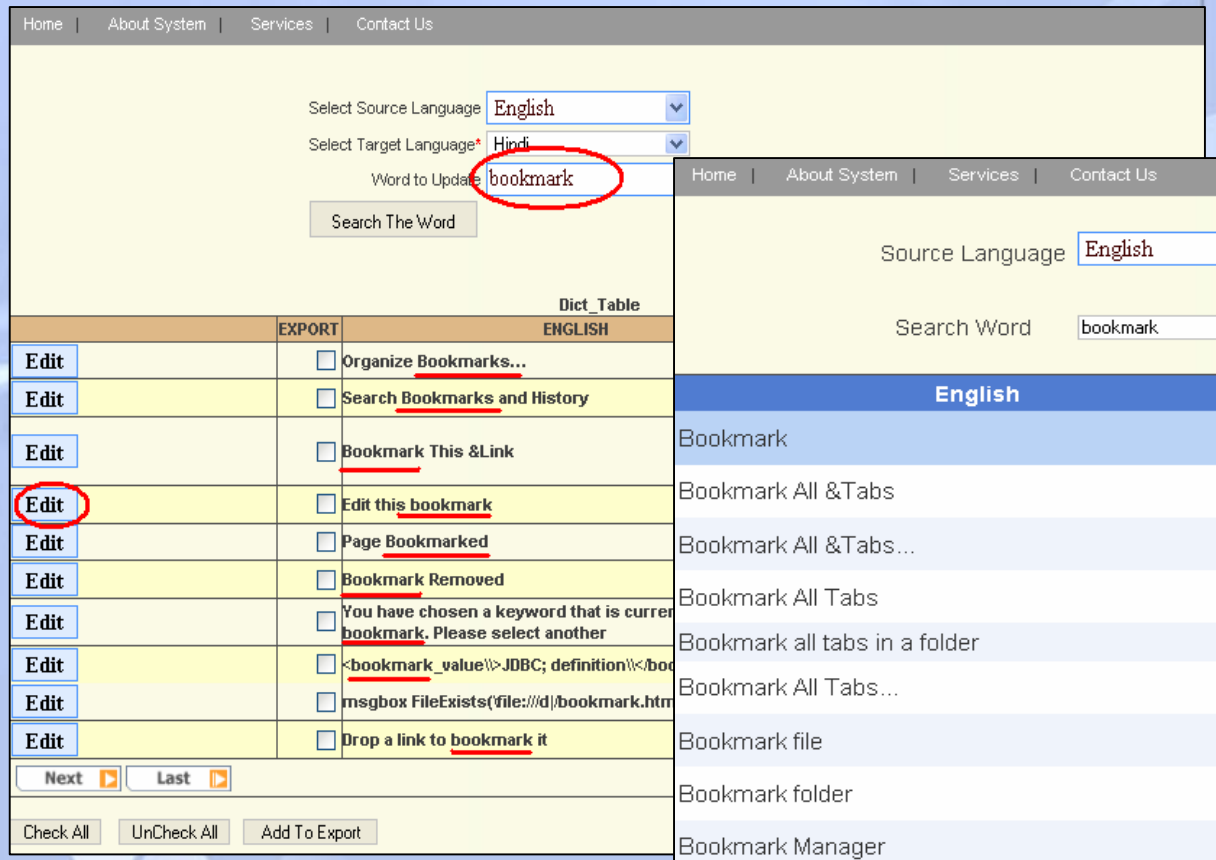


Localized Indian Language contents


Third Party tools / technologies
• Machine Translation
• Term banks
• Spellchecker
• Translation memories …
Web Services
Research outcome & 3rd party utilization and feedback provisioning
LPMF/Language Switch
Glossary, Term banks, Translation memories, …
C-DAC tools / technologies
Work flow / Configuration management
TDIL Research Outcome
• Machine Translation, CLIA.
• OCR
• Language resources
• spellchecker and NLP tools
Web Services
End User Experience, Term / translation, validation Knowledge base …

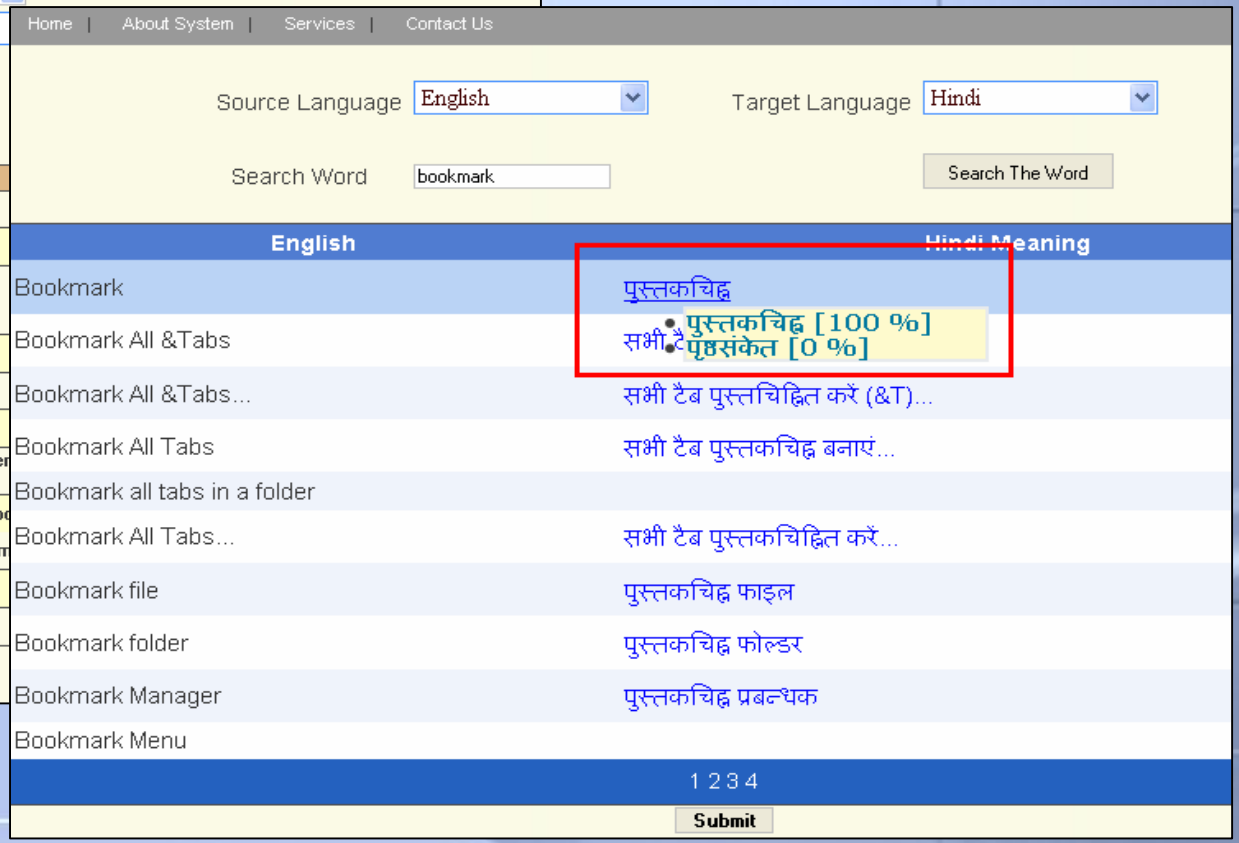
Community participation - Improving the Translations

Validating and improving the translations of NROP I using public voting system



Search word & Edit

Word Rating



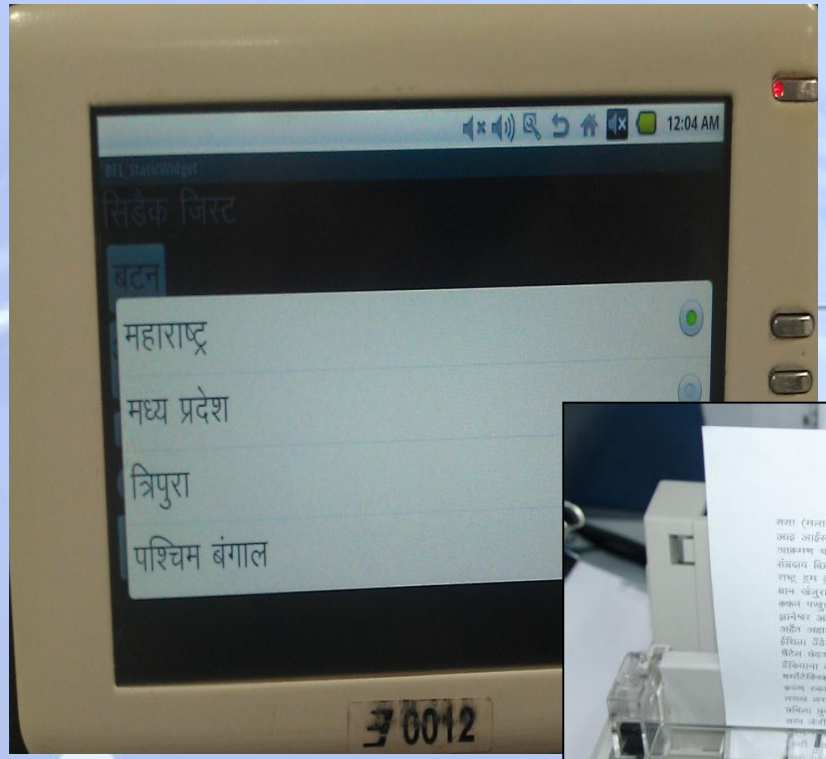
Localization of most of the popular FOSS

Localization of most of the popular FOSS in all the 22 Indian languages, with their alternate script:

- Libre / Open Office which is an entire office suite and includes
- ThunderBird (E-mail client)
- Firefox (Web-Browser)
- GAIM (Multiprotocol messenger for chatting)
- Linux Operating system (BOSS – Bharat Operating System and Solutions)
- Joomla
- And many more

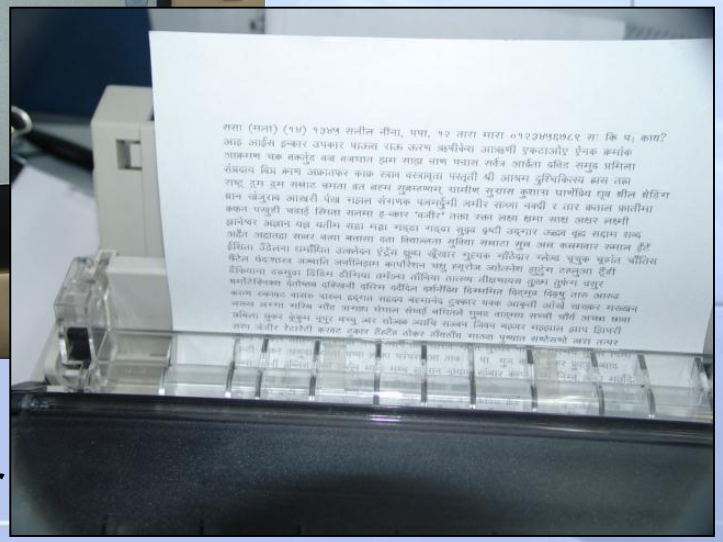
Localization - mobiles / PDAs

Indian language support on Android platform for use under MNREGA scheme



Indian language support to Mobiles

Dot / Line Matrix Printer



Conclusions

- Standards adherence is crucial (especially Unicode, INSCRIPT keyboard, Open Font Format, Script Grammar-Sort order, GIGW, Localization Guidelines, CLDR, FUEL, W3C and WCAG, Term banks)
- Compliance testing / Validation as per the guidelines / best practices.
- Localization Project management framework for SIM release and updates of e-governance applications.
- Use of content management system at backend is crucial.
- MMP specific term bank creation and sharing with developers.

Localization Guidelines
Minimal Set

Guidelines

Point 1: Default Homepage in Local Language

Point 2: 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.

Point 3: All Menu titles of the web pages in Local Language.

Point 4: All Web-pages should be developed using UTF-8 encoding.

Point 5: Lang attributes lang="mr" should be specified (replace “mr” with local language).

Point 6: Meta tags should be defined in Local Language.

Point 7: Image ALT/Captions, titles and text should be in Local Language.

Point 8: Font for the website should be provided through latest HTML5 Font SRC provisions.

Guidelines

Point 9: All Downloads (PDF, DOC, Excel, etc.) should be in Local Language with Unicode compliance.

Point 10: All page titles in Local Language.

Point 11: Numbers should be in Local Language on all pages, Documents, PDFs, Images, etc.

Point 12: Contact us information given in Local Language.

Point 13: All user defined alert/error/pop-up messages are in Local Language.

Point 14: Feedback form should be available in Local Language.

Point 15: Typing facility in Local Language should be given for Interactive Website.

Point 16: Typing should support INSCRIPT layout.

Guidelines

Point 17: Onscreen Floating keyboard should be made available.

Point 18: Provision for increasing font size is available.

Point 19: In-site search support is available for Local Language .

Point 20: Website works on Hand held devices.

Point 21: Site map of website is in Local Language.

धन्यवाद
Nurturing living languages
जीवित भाषाओं का पोषण