

The Indian Experience in Multilingual Computing

**S. P. Mudur &
Durgesh Rao
NCST, India**



राष्ट्रीय साँफ्टवेयर
प्रौद्योगिकी केन्द्र



Background



Indian Multilingual Context:

- 18 official languages
- 10 official scripts
- Millions of native speakers in each
- Low literacy and English literacy

**Huge potential for
multilingualisation and
localisation**



Definitions



Multilingualisation:

The process of designing, structuring and developing a system or product so that it has the capability to simultaneously handle multiple languages.



Definitions

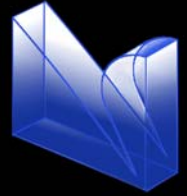


Localisation:

The process of modifying (adapting) a system or product to be easily used in the local environment, language and culture of a particular region or locality.



Definitions



Internationalisation:

The process of designing, structuring and developing a system or product so that it is region independent and can be easily localised to be used in a new region or locality.



Definitions



Globalisation:

The combined process of internationalisation and localisation of a system or product to address the global market (use).



Software Localisation Steps



- **Language Enabling**
 - encoding
 - rendering
 - input, cursor & edit key controls
- **User Interface**
 - Menus, dialogues, messages, documentation, ...
 - Icons, layouts, colour scheme, ...
 - Interaction modes
- **Environmental**
 - operational/documentation standards/practices
 - interoperability issues
- **Cultural**
 - human behaviour models, rules, ...



Software Internationalisation

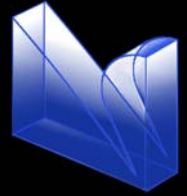


The Current Process:

- **Language Enabling**
 - Unicode encoding
 - OpenType rendering support
 - Keyboard mapping (cursor, edit key control ?)
- **User Interface**
 - Menus, dialogues, messages, ... (Resource files ?)
 - Documentation, online help, tutorials, ... ?
 - Icons (Resource files ?), layouts, colours ?
 - Interaction modes ?



Software Internationalisation

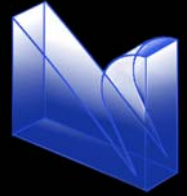


The Current Process: (contd.)

- **Environmental**
 - operational/documentation standards/practices (locale – too simplistic?)
 - interoperability issues (cut and paste, plug and play, conversion to national standards, ...)
- **Cultural ?**



Localisation for Indian Regions



- **Localisation of printed material began in late 18th century and have continued to date.**
 - **Hotmetal, Linotype, Monotype, Phototypesetting and Digital typesetting.**
- **Computer related localisation efforts began in early 70s**
 - **RMK Sinha & HN Mahabala, IIT Kanpur - Card punching scheme**
 - **HLMN Narasimham, CMC Ltd - Display terminal**
 - **SP Mudur, LS Wakankar, TIFR – coding, keyboard, shaping, display, hard copy etc.**
 - **Many other efforts**



Localisation for Indian Regions



- **Early 80s**

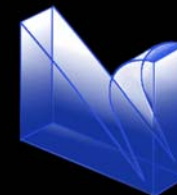
- **RMK Sinha, Mohan Tambe, IIT Kanpur – GIST terminal**
- **DoE - ISCII encoding, common encoding for all Indic scripts, modal in nature; TDIL programme**
- **PK Ghosh, RK Joshi, SP Mudur, NCST – Metafont, TeX**

- **Late 80s**

- **S. Ramani, NCST, Rajeev Sangal, IITK – Machine translation**
- **V. Bhatkar, M Tambe, CDAC – GIST Card**
- **SP Mudur, RK Joshi - Text processing, font design, ...**
- **Softek Ltd. – Terminals, Text processing tools, ...**
- **Modular Ltd., ITR - DTP fonts**
- **TCS, CMC, ... - Word processors, ...**
- **Many Others**



Vividha – Multilingual text processing tools



Vividha : Preview

Vividha File View Browse Options

1 2 3 4 5 6 7 8

1 **Script samples at 18 Point**

2 Devnagari: विविधा विविधा विविधा विविधा

3 Bengali: বিবিধা বিবিধা বিবিধা বিবিধা

4 Oriya: ବିବିଧା ବିବିଧା ବିବିଧା ବିବିଧା

5 Tamil: விவிதா விவிதா விவிதா விவிதா

* File I:\VIVIDHA2\VTX\ALL_SAMP.VTX Page 1 Left 0.00 Top 0.00
F1 Help F6 Editor {Browse- F5 Greek F7 Prev F8 Next} F9 Menu F0 Script





Vinyas – Calligraphic font design

The image displays a calligraphic font design on a grid. The main focus is the letters 'S' and 'T'. Below the letters is a ruler with 8 columns, each containing a diagram of a letter's stroke order. The diagrams are numbered 40 through 57. The letters 'S' and 'T' are shown in a large, bold, black font. The grid is composed of a fine dotted pattern and a coarser grid of vertical lines. The ruler below the letters is marked with numbers 40 through 57, corresponding to the stroke order diagrams. The diagrams show the path of the pen for each letter, with arrows indicating the direction of the stroke. The 'S' diagram shows a continuous curve, while the 'T' diagram shows a vertical stem and a horizontal crossbar.

40	41	42	43	44	45	46	47
50	51	52	53	54	55	56	57



Localisation for Indian Regions



- **90s**

- **Unicode based on ISCII 88**
- **Local language Products in the Market**
 - CDAC, Modular, Softek, Applesoft, Aakruti, Webdunia, Mithi, RK associates, ...

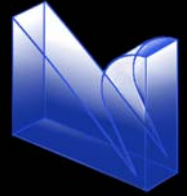
Late 90s

- **Localisation Efforts by Microsoft**
 - Language Enabling
 - Locale Definition
 - Word 2000 localised to some extent
- **Localisation Efforts by IBM**
 - Unicode based software components (ICU)
 - Locale definitions



Lotus Notes localised to some extent

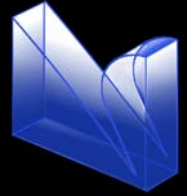
Why have there not been major localisation efforts for Indian regions?



- **Non-availability of internationalised software**
- **Not all 8-bit clean software implementations**
- **Lack of adequate market size**
- **Indian official language policy – trilingual**
 - **State, National and English languages**
- **Very special shaping needs of Indian scripts**
 - **During large frame computer days, shaping was done in hardware (Character ROMs)**
 - **On desktops, shaping is done in software, usually the graphics device interface (GDI)**
 - **Non-linear shaping rules not supported.**
- **Lack of standards and adherence to standards**



Why have there not been major localisation efforts for Indian regions?



Without shaping support at GDI level, not possible to enable ISCII (or Unicode) encoding during localisation.

The Result:

- Language enabling carried out through a linguistically incorrect font based solution.
- Each new developer introduced her/his own graphic break-up of the symbol set (incomplete).
- Over 50 different encodings (glyph level) and keyboard mappings in use today.
- Virtually no interoperability.



Local language software development



Some success has been achieved in local language software application development.

Examples include word processors, publishing packages, data base applications, spell checkers, financial applications, land records, transliteration packages, and a variety of turn-key projects.

Considerable Local language work for the web:

Until recently browsers did not have support for Unicode. No Indic script shaping support either. As a result web applications either used glyph level encodings with downloadable or dynamic fonts or special plug-ins for display.

Again no interoperability.



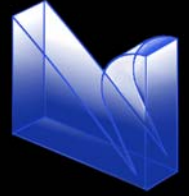
NCST's recent efforts in Windows localisation



- Windows NT 5 (Windows 2000)
- **Language enabling**
 - Common Unicode shaping engine for all left to right Indic scripts
 - One OpenType font each, for 12 scripts
 - Inscript Keyboard drivers for each script
 - Locale definitions for all official languages of India
 - Sort procedure for Hindi and Tamil
 - Edit controls



Open Type Fonts for Windows 2000™



Devanagari Consonants

क	ख	ग	घ	ङ	
च	छ	ज	झ	ञ	
ट	ठ	ड	ढ	ण	
त	थ	द	ध	न	
प	फ	ब	भ	म	
य	र	ल	व	श	ष
स	ह	ळ	क्ष	ज्ञ	

Devanagari Varnamala

अ आ इ ई उ ऊ ऋ ॠ ऌ ॡ ऋ ॠ ऌ ॡ

DEVANAGARI VARNAMALA



Localisation in Office XP by Microsoft



- Office XP builds on the language enabling work done earlier for Windows 2000.
- **All Office XP tools enabled for Indic scripts**
- **Supports 6 Indic scripts in current release.**
 - Devanagari, Tamil, Gujarati, Telugu, Kannada, Gurmukhi
- **More OpenType fonts for Devnagari.**
- **Spell checker for Hindi.**



NCST's recent efforts in Linux localisation



X windows level support for shaping Indic scripts.

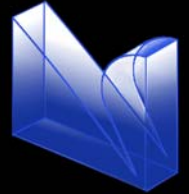
Use of Unicode standard for handling multi-lingual text.

No need to change Search engine and/or Sorting algorithm as we use UTF-8 transformation format of Unicode.

OpenType font support using FreeType library.



Linux Localisation Efforts at NCST



The screenshot shows a file manager window titled `/home/keyur/test`. The menu bar includes `File`, `Edit`, `Settings`, `Layout`, `Commands`, `Desktop`, and `Help`. The toolbar contains icons for `Back`, `Up`, `Forward`, `Rescan`, `Home`, `Icons`, `Brief`, `Detailed` (selected), and `Custom`. The location bar shows `/home/keyur/test`.

The left sidebar shows a tree view of directories, with `test` selected. The main pane displays a table of files and folders:

Name	Size	MTime
..	4096	Oct 17 15:29
रामायण	39313	Oct 17 16:26
महाभारत	464530	Oct 17 16:25
पगार	419	Jul 28 17:41
डिपार्टमेन्ट ओफ ईलेक्ट्रोनिक्स	164	Oct 17 16:23
aircut3.ttf	21548	Feb 29 2000
Bदुर्गशकुमार	356	Oct 13 14:54
benjamingothic.ttf	19096	Feb 29 2000
* compile	336	May 9 14:44

The status bar at the bottom indicates `39,313 bytes in 1 file` and includes a `Show all files` button.

Linux Localisation Efforts at NCST



प्री खेल #972996405

खेल निर्धारण सहायता

नया खेल फिर से बीज अनावृत्ति अंक विशेषता निकास

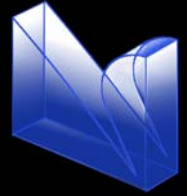
5♦	J♥	9♥	A♣	A♠	A♠	A♠	
4♠	7♠	8♠	A♥	Q♠	K♠	5♠	Q♠
K♠	J♠	4♠	9♥	2♠	0♠	4♠	3♠
6♥	9♠	A♥	2♠	8♥	6♠	10♥	6♠
2♠	10♠	2♥	6♠	6♠	6♠	J♥	Q♠
7♥	K♠	Q♥	7♥	6♦	3♣	3♣	3♥
2♥	K♠	6♦	2♥	9♦	2♣	2♣	2♥

प्रश्न

निकास करे?

हा ना

Linux Localisation Efforts at NCST



Mumbai University College List – Netscape 6 {Build ID: 2000092909}

File Edit View Search Go Bookmarks Tasks Help

http://202.141.151.139/college/collist.asp Search

Home My Netscape Mozilla.org Tinderbox Bonsai Bugzilla Open Directory

My Sidebar Tabs

What's Related

Search

Search Results

Bookmark Search Results

Tinderbox

Bookmarks

College List for the year (2000)

College Id	College Name (Englist/Devnagri)
1	Academy of Architecture अकॅडेमी ऑफ आर्किटेक्चर
2	Acharya Research Centre आचार्य रिसर्च सेंटर
3	Akbar Peerbhoy College of Commerce and Economics अकबर पीरभाय कॉलेज ऑफ कॉमर्स एन्ड इकनॉमिक्स
4	Aliyavar Jung National Institute for the Hearing Handicapped अलियावर जंग नॅशनल इन्स्टिट्यूट फार द हियरिंग हॅन्डिकॅपड
5	All India Institute of Diabetes ऑल इंडिया इन्स्टिट्यूट ऑफ डायबेटीस
6	All India Institute of Physical Medicine and Rehabilitation ऑल इंडिया इन्स्टिट्यूट ऑफ फिजिकल मेडिसिन एन्ड रिहॅबिलीटेशन
7	Ananthacharya Indological Research Institute अनन्ताचार्य इंडोलॉजिकल रिसर्च इन्स्टिट्यूट
8	Anjuman-I-Islam Urdu Research Institute

Document: Done

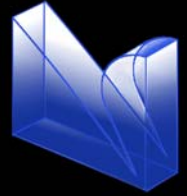
Business Tech Fun Interact

NCST's recent efforts in Open Office localisation



- OpenOffice.org Components
 - OO.o Word Processor (Star Writer)
 - OO.o Presentation Tool
 - OO.o Spread Sheet Tool
 - OO.o Drawing Tool
 - OO.o Charting Tool
 - HTML Document
 - Master Document

Some Screen Shots...



The screenshot shows the OpenOffice.org 627 presentation software interface. The title bar reads 'Untitled1 - OpenOffice.org 627'. The menu bar includes 'File', 'Edit', 'View', 'Insert', 'Format', 'Tools', 'Slide Show', 'Window', and 'Help'. The toolbar contains various icons for editing and presentation control. The main window displays a slide with the title 'अनुशासन' (Anushasan) in a large font. Below the title, there are three bullet points in Hindi:

- अनुशासन मनुष्य के व्यक्तित्व का परिचायक है।
- जीवन में अनुशासन का बहुत महत्व है।
- जीवन में अनुशासन मनुष्य को स्वच्छ बनाता है।

The status bar at the bottom shows 'TextEdit: Paragraph 1, Row 1, Column 8', '34.16 / -2.35', '23.91 x 3.51', '71%', '* Slide 1 / 1', and 'Home'.

Some Screen Shots...



The screenshot shows the OpenOffice.org 627 interface. The window title is "svc26.tmp.sxd - OpenOffice.org 627". The menu bar includes File, Edit, View, Insert, Format, Tools, Modify, Window, and Help. The toolbar contains various icons for editing and presentation. The Graphics Styles panel is open, showing "Continuous" fill, "0.00cm" size, "Black" color, "Invisible" line, and "Blue 7" style. The slide content includes a blue oval and a text box with the following text in Hindi:

पाकिस्थान के राष्ट्रपति जनरल परवेज मुशर्रेफ के इस कयन को अमरीका ने नही माना है कि रमजान के माह मे अफगानिस्थान पर हमले बन्द कर दिए जाने चाहिये।

The status bar at the bottom shows "Text Frame 'पाकिस्था...' selected", coordinates "11.64 / 11.95", size "0.00 x 0.00", zoom "34%", and "Slide 1 / 1".

Some Screen Shots...



The screenshot displays the OpenOffice.org 627 interface. The main window shows a spreadsheet with the following data:

	A	B
1	अजय	ajey@yahoo.com
2	मनोरंजन	manu@ncb.ernet.in
3	भूपेश	bhupesh@yahoo.co
4	शिखा	shikha@ncb.ernet.in
5	पवन	pawan@rediffmail.c
6	अमिठ	amit@yahoo.com
7	श्रीका	shikha@ncb.ernet.in

The 'Save As' dialog box is open, showing the file path C:\OpenOffice60\user\backup. The file list contains one entry:

Title	Size	Modified Date
hello.sxc	6092 Bytes	10/25/2001, 12:43:26 P

The dialog box also shows the file name 'पहला' and file type 'Calc 6.0'. The 'Automatic file name extension' checkbox is checked, and the 'Save with password' checkbox is unchecked.

User Interface Localisation



No significant efforts have been made so far.

Some feeble attempts by small groups around the country.

Examples:

- **TDIL programmes at JNU and Banasthali Vidyapeeth**
- **BARC**
- **Indian Linux Users group**
- **MAIT supported work**

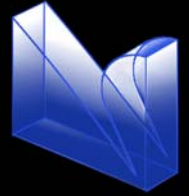
Terminology yet to be standardised.

Documentation, Online help, Tutorials,... None

Feeble attempts at speech synthesis.



User Interface Localisation



**Some efforts in OCR.
Handwritten character recognition?**

**Considerable efforts in Machine Translation and
some in Lexical Resource creation.**

**Much more needed, if these efforts have to
provide assistance in translation.**



Environment Related Localisation



Locales defined for all major languages during NCST- Microsoft collaboration.

Also being used by IBM.

There are built-in functions for handling locale parameters.

Sort procedures defined: Hindi and Tamil only.

Now same facilities in Linux also.



Culture Related Localisation



No Attempts so far.



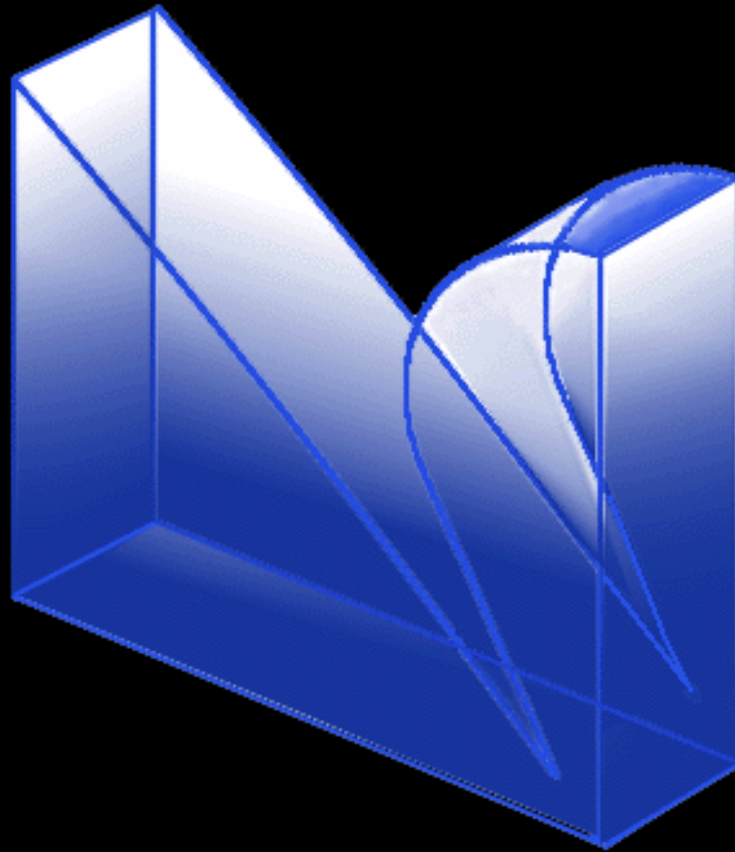
Localisation Engineering



Most groups are yet to become aware of this area.

Indian companies have not yet realised the potential market opportunities in this domain.





Thank You!