

Sinhala Unicode Developer Workshop

Muthu Nedumaran
(muthu@murasu.com)

What is Unicode?

- Universal Character Set
 - All of the major scripts
 - Simple and consistent manner
 - Alphabetic, syllabic and ideographic scripts
- Version 4.0
 - 50,000 characters
 - Over 90 scripts

Unicode Implementation

- All major operating systems
 - Windows, MacOS, Linux, PalmOS, WinCE, Symbian
- WWW
 - HTML 4.0, XML, Java, JavaScript
- Applications
 - MS Office, OpenOffice, InDesign, Acrobat, IE and many more

Inside a Unicode Sinhala Font

Glyphs (glyf)	OpenType accepts glyphs in TrueType or Type1 format
Character to Glyph Mapping Table (cmap)	Maps character codes to glyphs. Straight one to one mapping. For Indic (& Hebrew, Arabic etc) scripts, number of glyphs required are more than number of characters defined
OpenType Tables	GSUB table provides substitution information. GPOS table provides positioning information. Can be used to minimise the number of glyphs required and thus the size of a font

Inside a Unicode Text Document

- Unicode Marker (Text)
 - Byte ordering dependant
- Characters “Only”
- No Ligatures or “Unencoded” shapes
- No font information
 - Text is not bound to a font
- Sinhala and Tamil recognised respectively

Input Method Editors

- Legacy Keyboard Drivers
 - Mapped to ASCII
 - Mapped to 8bit
- Sinhala Unicode IME’s
 - Vowels, Consonants, Ligatures
 - Key Layouts
 - FontTester

Unicode Friendly Applications

- Currently Supported:
 - Text Editors/Word Processors
 - Browsers
 - Databases
- Possible Expansion:
 - Spell Check/Dictionary
 - Client (Desktop) Applications
 - Other utilities and tools

DEMO

- Sinhala Font and Text
- Legacy Text (7bit Font)
 - ඊ ඉ ඊ ඉ
- Unicode Text (Unicode Font)

Unicode Filenames

- Windows
- Mac OS X

BREAK

Unicode Text Format

- ANSI, ASCII, UTF-8, UTF-16
- Windows Notepad
- Email Messages
- HTML Documents
- RTF Format

Unicode Strings and APIs

- Windows
- MacOS
- Java
- JavaScript (>1,3)
- PHP

Parsing Strings

- Determining if text is Unicode
- Determining Consonants, Vowels, Marks etc
- How do I know if the text is Unicode?
- Byte-Stripping
- Searching

Demos

- External Rendering vs Internal Representation
 - FontTesterTool
- Handling Unicode Strings
- Converting Legacy Strings

Unicode APIs

- WideString
 - Functions
- Messages
- ANSI vs Unicode

A Simple Unicode Application

- English, Sinhala and Tamil on the same document
- Display messages in Sinhala/Tamil
- Text input in Sinhala/Tamil

DEMO:

- A Simple Unicode Application

Unicode Web Applications

- HTML and JavaScript
- Header
- Embedding Fonts
- Text strings

Unicode Web Applications

- Forms and Fields
- User Input
- IME Handling

Server Side

- Database Support
- Manipulating Strings
- Co-existence
 - Traditional/Legacy Text