

Sinhala Unicode Developer Workshop

Muthu Nedumaran
(muthu@murasu.com)

What is Unicode?

- Computers deal with numbers
- Characters are assigned numbers
- Before Unicode, these numbers were assigned in many different ways
 - Encoding, Code-Pages
 - “Code-point” overloading!
 - No way to determine what the character is
 - Data Corruption

What is Unicode?

- Unique number for every character
 - no matter what the platform
 - no matter what the program
 - no matter what the language
- Easy to determine what that character is and which language it is from

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

Benefits of Unicode

- Can be incorporated into
 - Desktop Apps
 - Client-Server Apps
 - Multi-tiered / Web Apps
- Significant cost savings
- Single implementation targeted across:
 - Multiple Platforms
 - Multiple Languages, countries
 - No re-engineering
- Data to be transported through many different systems
 - No data corruption!

Sinhala Unicode

- Only “Characters” are encoded
 - All Indic family of scripts, including Tamil
- All rendering information are in the font and shaping mechanism in the OS platform
- Applications do not have to deal with “ligatures” or “conjuncts”
- Text represented as WideStrings

Unicode Implementation

● All major operating systems

- Windows, MacOS, Linux, PalmOS, WinCE, Symbian

● WWW

- HTML 4.0, XML, Java, JavaScript

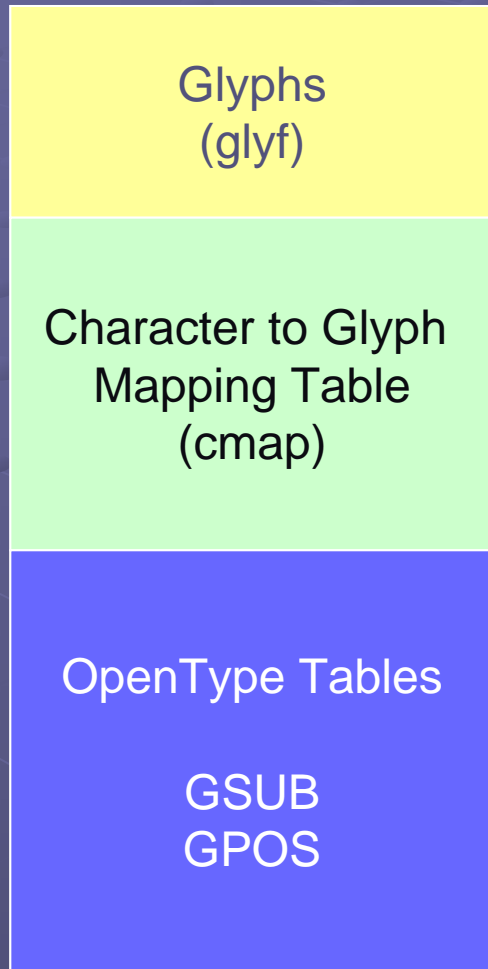
● Applications

- MS Office, OpenOffice, InDesign, Acrobat, IE and many more

Sinhala Unicode

● Character Block

Inside a Sinhala Unicode Font



OpenType accepts glyphs in TrueType or Type1 format

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

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
 - Handles Vowels, Consonants & Signs only
 - Key Layouts
 - FontTester

Unicode Friendly Applications

● Currently Supported:

- Text Editors/Word Processors
- Web Servers, Browsers
- Development Tools, Databases

● Possible Expansion for Sinhala:

- Spell Check/Dictionary
- Client (Desktop) Applications
- Other utilities and tools

DEMO

- Sinhala Font and Text
- Legacy Text (7bit Font)
 - ලස්සන මගෙ රට
- Unicode Text (Unicode Font)
 - ලස්සන මගෙ රට

Unicode Filenames

- Windows
- Mac OS X

BREAK

5 June, 2004

© Murasu Systems Sdn. Bhd., Malaysia

Unicode Text Format

- ANSI, 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

Converting Legacy Strings

- Base characters
- Post modifiers
- Pre modifiers
- Two-part modifiers
- Half-glyphs
- Special symbols

Unicode APIs

- WideStrings
 - 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

- Dynamic Fonts
 - EOT Font Format
 - Microsoft WEFT
- Forms and Fields
- User Input
- IME Handling

Server Side

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