



Locale, IBM ICU and CLDR

Subir B. Pradhanang
Development Engineer
Madan Puraskar Pustakalaya, Nepal
URL: www.mpp.org.np
Email: subir@mpp.org.np



Contents:

- Basics of Locale
 - What is a Locale?
 - Locale Naming
 - Locale Definition File
 - Building a locale

- ICU

- Features of ICU

- CLDR
 - Introduction
 - Goals
 - Brief History

- Releases
 - Latest – CLDR 1.3
 - Previous – CLDR 1.2



Basics of Locale

→ What is a Locale?

- . Locale can be referred to as a collection of information associated with a country or region
- . Includes Language spoken in the region, Scripts, Dates/time formats, Number/Currency formats, Measurement System, Collation (for sorting, searching), etc.
- . Part of the GNU C library (glibc) package
- . Every language must have it's own locale
- . Many localized software are dependent on locale eg. Gnome desktop, Sort utilities, etc.
- . Nepali locale developed and submitted by MPP and named as ne_NP
- . Locale built using locale definition file and charmap file



Basics of Locale

Contd...

→ Locale Naming:

- . Locale is described by the language, country and character set
- . Naming convention for a locale:

`lang_territory.codeset[@modifiers]`

Here,

- . lang = 2-letter language code defined in ISO 639:1988[3], 3-letter language code is defined in ISO 639-2[4] which is used in the absence of the 2-letter version
- . territory = 2-letter country code defined in ISO 3166-1:1997[6]
- . codeset = Character set used
- . modifiers = Optional; adds more information to the locale by setting options

Eg. `fr_CA.ISO-8859-1` means French language spoken in Canada using ISO-8859-1 character set



Basics of Locale

Contd...

→ **Locale definition file's contents and definition:**

- . LC_CTYPE - Category related to encodings
- . LC_COLLATE - Category related to sorting
- . LC_MESSAGES - Category related to the language for messages the software outputs
- . LC_MONETARY - Format to show monetary numbers, Currency symbol, comma or period
- . LC_NUMERIC - Number format to show position of decimal digit separators
- . LC_TIME - Category related to format to show time and date, such as name of months and weeks, order of date, month, and year, and so on.

- . New categories introduced:
 - . LC_PAPER – paper size
 - . LC_NAME – personal name format
 - . LC_ADDRESS – address format
 - . LC_TELEPHONE – telephone number
 - . LC_MEASUREMENT – measurement units
 - . LC_VERSION – locale version



Basics of Locale

contd....

→ **Building a locale:**

- Two things needed for building a locale:
 - locale definition file
 - charmap file*
- In our case, locale definition file is ne_NP and charmap file is UTF-8
- Compile them to create a locale using localedef command:
Syntax: localedef [-f {charmap}] [-i {input} {name}]

Eg. for Nepali locale:

```
# localedef -f UTF-8 -i ne_NP ne_NP
```

where, -f option specifies character map source file

-i option specifies locale definition source file

NOTE: If you are defining your own character set, then you need to create a charmap file for it giving every character a symbolic name and describing the encoded byte strings.



ICU

- . Acronym for International Components for Unicode
- . Mature, widely used set of C/C++ and Java libraries for Unicode support, software internationalization and globalization (i18n/g11n)
- . widely portable and gives applications the same results on all platforms and between C/C++ and Java software
- . ICU gives rendering support to:
 - OpenOffice.org
 - Sun's and IBM's java distributions
- . ICU currently supports several Indic languages, but there are still some issues regarding Nepali rendering in OpenOffice.org



ICU Features

- . ICU libraries provide robust and full-featured Unicode services on a wide variety of platforms, without sacrificing performance
- . supports the most current version of the Unicode standard
- . offers great flexibility to extend and customize the supplied services, which include:
 - Text: Unicode text handling
 - Comparison: Language sensitive collation and searching
 - Locales: Comprehensive locale data (230+) and resource bundle architecture
 - Complex Text Layout: Arabic, Hebrew, Indic and Thai
 - Time: Multi-calendar and time zone
 - Formatting and Parsing: dates, times, numbers, currencies, messages and rule based
- . ICU is an open source development project sponsored, supported, and used by IBM
- . dedicated to providing robust, full-featured, commercial quality, freely available Unicode-based technologies



CLDR

→ Introduction

- . Acronym for Common Locale Data Repository
- . Relatively new project: 2004
- . A central location for locale data managed by Unicode Consortium that is becoming a central reference location for all locale data
<http://www.unicode.org/cldr>
- . By far the largest standard repository of locale data

→ Goals

- . Common, necessary software locale data for world language
- . general XML format for the exchange of locale information
- . Freely available



CLDR

Contd...

→ Brief History

- . First CLDR version under the sponsorship of the Unicode Consortium was version 1.1
- . The CLDR project originally developed under the sponsorship of the Linux Application Development Environment (aka LADE) Workgroup of the Free Standards Group's OpenI18N team
- . CLDR 1.0 approved in Jan, 2004 by the OpenI18N steering committee; Founding members of the workgroup consisted of IBM, Sun and OpenOffice.org



Latest Release: CLDR 1.3

- . Released: June, 2005
- . Contains data for:
296 locales: 96 languages and 130 territories
- . For the first time in CLDR, POSIX formatted data also available

Previous Release: CLDR 1.2

- . Released: November, 2004
- . Contains data for:
Approved: 232 locales: 72 languages and 108 territories
Draft: 63 locales: 27 languages and 28 territories