Software Globalization and Adding Languages on Computers and Mobile Devices

Craig Cummings
Unicde Technical Committee Vice-Chair
ANSI INCITS L2 Committee Chair
Unicode Emoji Subcommittee Animal/Science Category Manager
VMware Globalization Team Lead
Why Globalize Software?

IDC Worldwide Black Book:
IT spend by language

<table>
<thead>
<tr>
<th>Language</th>
<th>Number of countries in which language plays a significant role</th>
<th>Share of IT spend</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>54</td>
<td>42.28%</td>
</tr>
<tr>
<td>Japanese</td>
<td>1</td>
<td>8.87%</td>
</tr>
<tr>
<td>Simplified Chinese</td>
<td>2</td>
<td>8.47%</td>
</tr>
<tr>
<td>German</td>
<td>4</td>
<td>4.77%</td>
</tr>
<tr>
<td>Spanish</td>
<td>24</td>
<td>3.82%</td>
</tr>
<tr>
<td>French</td>
<td>31</td>
<td>3.60%</td>
</tr>
<tr>
<td>Portuguese</td>
<td>5</td>
<td>2.95%</td>
</tr>
<tr>
<td>Korean</td>
<td>1</td>
<td>1.86%</td>
</tr>
<tr>
<td>Italian</td>
<td>2</td>
<td>1.79%</td>
</tr>
<tr>
<td>Russian</td>
<td>11</td>
<td>1.72%</td>
</tr>
<tr>
<td>Dutch</td>
<td>5</td>
<td>1.37%</td>
</tr>
<tr>
<td>Traditional Chinese</td>
<td>3</td>
<td>1.50%</td>
</tr>
<tr>
<td>Arabic</td>
<td>25</td>
<td>.95%</td>
</tr>
</tbody>
</table>
Why Globalize Software? -- English vs. Localized Page Views

Japan
- Japanese: 96%
- English: 4%

Korea
- Korean: 95%
- English: 5%

China Mainland
- Simplified Chinese: 86%
- English: 13%
- Others: 1%

Taiwan
- Traditional Chinese: 82%
- English: 18%

France
- French: 79%
- English: 21%

Germany
- English: 28%
- German: 72%
Globalization (G11n): The complete process of making your application available in multiple languages consisting of 2 sub processes Localization and Internationalization.

Internationalization (i18n): Architecting and coding for language/locale independence; allows localization for target audiences that vary in culture, region, or language.

Localization (L10n): Adapting to specific languages/locales by translating text and adding locale-specific formatting to images and style sheets.

Translation (T9n): Converting the meaning of text in one language into another.

Localization Rocket Science – Machine Translation (MT)

- Rule-based MT
- Statistic MT
- Neural MT

Prevailing since late 2016
Dominating 10+ years until 2016

java.io.InputStreamReader
(use the constructor that includes an explicit encoding)

Code this

```
new InputStreamReader(outputStream, explicitEncoding);
```

Not this

```
new InputStreamReader(outputStream);
```
Internationalization Eye Chart

• Encodings, I/O, character support
  • Regular expressions including Unicode regular expressions
    • Because [A-Za-z0-9] works only for ASCII
  • Internationalized Domain Names (IDNs)/Internationalized Resource Identifiers (IRIs)
  • Supplementary character support

• Locales (BCP-47 compliant)
  • Determination
  • Representation
  • Negotiation
  • Fallback

• Locale-sensitive APIs
  • Externalization
    • Resource files and messages

• Formatting
  • Dates, times, time zones and calendars
  • Numbers
  • Currencies
  • Honorifics
  • Addresses and phone numbers

• Linguistic (not binary) Sorting
• Wrapping/boundary analysis (words, lines, sentences, etc.)

• Keyboard support
• User Interface (including bidi), fonts, and layout support
• Search, normalization, folding, etc.
• Security
• ...and more
You Probably Know Unicode for...
Hoping you know Unicode for writing systems
Did you know Unicode for this, too?
The Three (minimal) Requirements To Get Your Language On Computers and Mobile Devices

1. Fonts (already there for Latin)
2. Keyboards (layouts now a part of CLDR)
3. Unicode CLDR data (the short straw)

CLDR = Common Locale Data Repository
CLDR - Unicode Common Locale Data Repository

Core Data for New Locales

This document describes the minimal data needed for a new locale. There are two kinds of data:

1. Core XML Data - This is data that the CLDR committee needs from the proposer before a new locale is added. The proposer is expected to also get a Survey Tool account and contribute towards the Minimal Data.
2. Minimal Data Commitment - Data that is expected to be provided for each locale. If it is not supplied in a timely fashion, the committee may remove the locale.

(The parenthesis at the start of each line below has the approximate number of strings for each item.)

Core XML Data

First, make sure you have correct language code according to Picking the Right Language Identifier. Then collect the following data. Consider using the Core Data Submission Form to submit this data.

Note to translators: If you are having difficulties or questions about the following data, please contact us. Post a follow-up to your existing bug, file a new bug, or reply to the mailing list.

1. (04) Exemplar sets: main, auxiliary, index. [mainXXX.xml]
2. (02) Orientation [bidi writing systems only] in [mainXXX.xml]
3. (01) Plural rules [supplemental/plurals.xml]
4. (01) Default content script and region (normally, normally country with largest population using that language, and normal script for that). [supplemental/supplementalMetadata.xml]
5. (N) Verify the country data (i.e. which territories in which the language is spoken enough to create a locale) [supplemental/supplementalData.xml]
6. (N) Casing information (cased scripts only, according to ScriptMetadata.txt)
7. (N) Collation rules (non-Survey Tool)
8. (N) The result will be a file like: common/collation/xxx-lr.xml or common/collation/xxx-d.xml.
9. (N) Note that the "search" collators (which tend to be large) are not needed initially.

Recommended Core Data

The following are not required, but are strongly recommended:

1. (04) Exemplar set: punctuation. [mainXXX.xml]
2. (01) Ordinal rules [supplemental/ordinals.xml]
3. (N) Romanization table (non-Latin writing systems only) [spreadsheet, will translate into transforms/xxx-en.xml]
4. (N) If a spreadsheet, for each letter (or sequence) in the exemplars, what is the corresponding Latin letter (or sequence).
5. (N) More sophisticated users can do a better job, supplying a file of rules like transforms/Arabic-Latin-BGN.xml.

Minimal Data Commitment

This data is to be entered using the Survey Tool except as noted.

1. (44+) 4 main Date/Time formats, 12 long & abbreviated, format & stand-alone month-names, 7 long & abbreviated day-names, 2 long day periods.
2. (01) Name of the language in the language.
3. (N) For any country locales, name of the country in the language, name/symbol for that country’s currency. Must be at least one, for the default content locale.
4. (05) Date/time pattern, intervalFormatFallback
5. (00) (for Latin) decimal and grouping separators, decimal, currency, percent formats
6. (N) Names of countries (territories) with that language as official.
7. (N) Names of exemplar/Cities in multizone countries with that language as official.
8. (05) Timezone patterns [http://cldr.unicode.org/translation/tzzone]
10. (03) key names
11. (14) long/short unit names (time intervals)
The First Seven Steps to ’Core XML Data’

1. Exemplar characters (dictionaries, grammar, references)
2. Left-to-right, Right-to-left, other directions
3. Plural rules (one duck, two ducks, etc.)
4. Script/region data (ISO 15924, 3166, 639)
5. More region data
6. Casing (titlecase for Latin-based languages)
7. Collation (fancy word for sorting; order with exemplars)
Getting To A Better CLDR Interface
Getting to Exemplars: Step 1A - Main Letters
Getting to Exemplars: Step 1B - Index Letters
The Still Easy ‘Minimal’ Remainder

• Just Over 50 additional easy-to-find translations
  • (38) The brunt are some 12 month names and 7 day names
    • Translations often in dictionaries, apps, or other good references
  • (1) Name of the language in that language
  • (7) “Year”, “Month”, “Week”, “Day”, “Hour”, “Minute”, “Second”
  • (6) Next slide

• With caveats
  • Some translations may require a bit more ‘wordsmithing’
  • Abbreviations may not be a concept in a given language
The Last Six Strings To Minimal

- “Standard Time”
- “Daylight Time”
- “Gregorian Calendar”
- “Standard Sort Order”
- “Western Digits”
- <date> “at” <time>
# Months of the Year

## Lakota / Date & Time / Gregorian

<table>
<thead>
<tr>
<th>English</th>
<th>Lakota</th>
<th>Status</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>January</td>
<td>✓</td>
<td>M01</td>
</tr>
<tr>
<td>Feb</td>
<td>February</td>
<td>✓</td>
<td>M02</td>
</tr>
<tr>
<td>Mar</td>
<td>March</td>
<td>✓</td>
<td>M03</td>
</tr>
<tr>
<td>Apr</td>
<td>April</td>
<td>✓</td>
<td>M04</td>
</tr>
<tr>
<td>May</td>
<td>May</td>
<td>✓</td>
<td>M05</td>
</tr>
<tr>
<td>Jun</td>
<td>June</td>
<td>✓</td>
<td>M06</td>
</tr>
<tr>
<td>Jul</td>
<td>July</td>
<td>✓</td>
<td>M07</td>
</tr>
<tr>
<td>Aug</td>
<td>August</td>
<td>✓</td>
<td>M08</td>
</tr>
<tr>
<td>Sep</td>
<td>September</td>
<td>✓</td>
<td>M09</td>
</tr>
<tr>
<td>Oct</td>
<td>October</td>
<td>✓</td>
<td>M10</td>
</tr>
<tr>
<td>Nov</td>
<td>November</td>
<td>✓</td>
<td>M11</td>
</tr>
<tr>
<td>Dec</td>
<td>December</td>
<td>✓</td>
<td>M12</td>
</tr>
</tbody>
</table>
### Days of the Week

<table>
<thead>
<tr>
<th>Day</th>
<th>English</th>
<th>Lakota</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sun</td>
<td>Sunday</td>
<td>Ṣuŋpẹtuwakȟáŋ</td>
</tr>
<tr>
<td>Mon</td>
<td>Monday</td>
<td>Ṣuŋpẹtuwaŋži</td>
</tr>
<tr>
<td>Tue</td>
<td>Tuesday</td>
<td>Ṣuŋpẹtuŋwaŋpa</td>
</tr>
<tr>
<td>Wed</td>
<td>Wednesday</td>
<td>Ṣuŋpẹtuŋyamni</td>
</tr>
<tr>
<td>Thu</td>
<td>Thursday</td>
<td>Ṣuŋpẹtuŋtopa</td>
</tr>
<tr>
<td>Fri</td>
<td>Friday</td>
<td>Ṣuŋpẹtuŋzaptak</td>
</tr>
<tr>
<td>Sat</td>
<td>Saturday</td>
<td>Owáŋgyužaŋapi</td>
</tr>
</tbody>
</table>
What Success Looks Like On A Device

Language & Region preferences control the language you see in menus and dialogs, and the formats of dates, times, and currencies.

Preferred languages:

GWY
Cherokee — Primary

Lakȟólíyapi
Lakota

ᐃᓄᒃᑎᑐᑦ
Inuktitut

Diné bizaad
Navajo

Guarani
Guarani

Region: United States
First day of week: Sunday
Calendar: Gregorian
Time format: 24-Hour Time
List sort order: Universal
A Success Story: Cherokee (in their own words)

https://youtu.be/EEEu8ufwW08?t=1498
How To Get Started

• File an new ticket with the ‘Core XML Data’ here:
  • http://unicode.org/cldr/trac
  • For reference, Osage, Muscogee/Creek, Chickasaw examples here:
    • https://unicode.org/cldr/trac/ticket/10721
    • https://unicode.org/cldr/trac/ticket/11424
    • https://unicode.org/cldr/trac/ticket/11983

• Ping me with ticket #
  • craig@unicode.org
Saving the World with Computing

Kathy Yelick
EECS Professor, U.C. Berkeley

Associate Laboratory Director for Computing Sciences and NERSC Director, Lawrence Berkeley National Laboratory
Q&A

craig@unicode.org
ccummings@vmware.com
Postscript

• CLDR Release Timing
  • Spring and fall of each year; fall is a bigger release
  • Calendar on CLDR home page

• Unencoded scripts require more work (e.g., Mi’kmaq)
  • Happy to guide towards a Unicode character encoding proposal

• Patience is required
  • Data may only take an hour to prepare, but keyboards, fonts, and system integrations will take longer.

• Full localizations of user interfaces, online help, and other documentation are a different discussion
  • Happy to have that discussion, too.