

International Language Environments Guide

- [Previous: Cultural Conventions](#)
- [Next: Keyboard Differences](#)

Using Locale Categories for Localization

The localization of a product should be done in consultation with native users in that target language or region. Certain styles and information styles and formats might seem perfectly obvious and universal to the developer, but to the user, these look either awkward, wrong, or even offensive. The following pages describe the elements that the Solaris operating environment allows you to control and specify so that you can successfully internationalize your product.

Time Formats

Table 1-1 shows some of the ways to write 11:59 P.M.

Table 1-1 International Time Formats

Locale	Format
Canadian	23:59
Finnish	23.59
German	23.59 Uhr

Locale	Format
Norwegian	Kl 23.59
Thai	11:59 PM
U.K.	11.59 PM

Time is represented by both a 12-hour clock and a 24-hour clock. The hour and minute separator can be either a colon (:) or a period (.).

Time zone splits occur between and within countries. Although a time zone can be described in terms of how many hours it is ahead of, or behind, Greenwich Mean Time (GMT), this number is not always an integer. For example, Nekeybfoundland is in a time zone that is half an hour different from the adjacent time zone.

Daylight Savings Time (DST) starts and ends on different dates that can vary from country to country.

Date Formats

Table 1-2 shows some of the date formats used around the world. Notice that even within a country, there can be variations.

Table 1-2 International Date Formats

Locale	Convention	Example
Canadian (English and French)	yyyy-mm-dd	1998-08-13

Locale	Convention	Example
Danish	yyyy-mm-dd	1999-08-24
Finnish	dd.mm.yyyy	13.08.1998
French	dd/mm/yyyy	13/08/1999
German	yyyy-mm-dd	1999-09-18
Italian	dd.mm.yy	13.08.98
Norwegian	dd.mm.yy	13.08.98
Spanish	dd-mm-yy	13-08-98
Swedish	yyyy-mm-dd	1998-08-13

Locale	Convention	Example
GB-English	dd/mm/yy	13/08/98
US-English	mm-dd-yy	08-13-98
Thai	dd/mm/yyyy	10/12/2009

Numbers

Decimal and Thousands Separators

Great Britain and the United States are two of the few places in the world that use a period to indicate the decimal place. Many other countries use a comma instead. The decimal separator is also called the **radix** character. Likewise, while the U.K. and U.S. use a comma to separate groups of thousands, many other countries use a period instead, and some countries separate thousands groups with a thin space. Table 1-3 shows some commonly used numeric formats.

Table 1-3 International Numeric Conventions

Locale	Large Number
Canadian (English and French)	4 294 967 295,000
Danish	4 294 967 295,000

Locale	Large Number
Finnish	4 294 967 295,000
French	4 294 967 295,000
German	4 294 967.295,000
Italian	4.294.967.295,000
Norwegian	4.294.967.295,000
Spanish	4.294.967.295,000
Swedish	4 294 967 295,000
GB-English	4,294,967,295.00

Locale	Large Number
US-English	4,294,967,295.00
Thai	4,294,967,295.00

Data files containing locale-specific formats are frequently misinterpreted when transferred to a system in a different locale. For example, a file containing numbers in a French format is not useful to a U.K.-specific program.

List Separators

There are no particular locale conventions that specify how to separate numbers in a list. They are sometimes comma-delimited in Great Britain and the U.S., but often spaces and semicolons are used.

Currency

Currency units and presentation order vary greatly around the world. [Table 1-4](#) shows monetary formats in some countries.

Table 1-4 International Monetary Conventions

Locale	Currency	Example
Canadian (English)	Dollar (\$)	\$1 234.56
Canadian (French)	Dollar (\$)	1 234.56\$

Locale	Currency	Example
Danish	Kroner (kr)	kr 1.234,56
Finnish	Markka (mk)	1.234,56 mk
French	Franc (F)	1 234,56 F
German	Deutsche Mark (DM)	DM 1.234,56
Italian	Lira (L)	L1.234,56
Japanese	Yen	41,234 Yen
Norwegian	Krone (kr)	kr 1.234,56
Spanish	Peseta (Pts)	1.234,56Pts

Locale	Currency	Example
Swedish	Krona (Kr)	1.234,56 kr
GB-English	Pound	31,234.56 pounds
US-English	Dollar (\$)	\$1,234.56
Thai	Baht	2539 Baht
Euro	EUR	400,00 €

Note -

Local and international symbols for currency can differ. For example, the designation for the French franc is "F" in France but this is often written as FRF' internationally to distinguish it from other francs, such as the Swiss franc or the Polynesian franc.

Note -

Euro locales are based on the ISO8859-15 character set. See ["European Localization"](#) for available locales.

Be aware also that a **converted** currency amount can take up more or less space than the original amount. To illustrate: \$1,000 can become L1.307.000.

Language Word and Letter Differences

Word Delimiters

In English, words are separated by a space character. In languages such as Chinese, Japanese and Thai, however, there is often no delimiter between words.

Sort Order

Sorting order for particular characters is not the same in all languages. For example, the character "ö" sorts with the ordinary "o" in Germany, but sorts separately in Sweden, where it is the last letter of the alphabet. In some languages, characters have weight to determine the priority of the character sequences. For example, in Thai, the Thai dictionary defines sorting through the sequences of characters that have different weights.

Character Sets

Number of Characters

While the English alphabet contains only 26 characters, some languages contain many more characters. Japanese, for example, can contain over 40,000 characters, Chinese even more.

Western European Alphabets

The alphabets of most western European countries are similar to the standard 26-character alphabet used in English-speaking countries, but there are often some additional basic characters, some marked (or accented) characters, and some ligatures.

Japanese Text

Japanese text is composed of three different scripts mixed together: Kanji ideographs derived from Chinese, and two phonetic scripts (or syllabaries), Hiragana and Katakana.

Although each character in Hiragana has an equivalent in Katakana, Hiragana is the most common script, with cursive rather than block-like letter forms. Kanji characters are used to write root words. Katakana is mostly used to represent "foreign" words--words "imported" from languages other than Japanese.

Kanji has tens of thousands of characters, but the number commonly used has been declining steadily over the years. Now only about 3500 are frequently used, although the average Japanese writer has a vocabulary of about 2000 Kanji characters. Nonetheless, computer systems must support more than 7000 because that is what the Japan Industry Standard (JIS) requires. In addition, there are about 170 Hiragana and Katakana characters. On average 55% of Japanese text is Hiragana, 35% Kanji, and 10% Katakana. Arabic numerals and Roman letters are also present in Japanese text.

Although it is possible to completely avoid the use of Kanji, most Japanese readers find text containing Kanji easier to understand.

Korean text can be written using a phonetic writing system called Hangeul. Hangeul has more than 11,000 characters, which consist of 19 consonants, 21 vowels, and an optional 27 consonants. About 3,000 Hangeul characters from the entire Hangeul characters are usually used in Korean computer systems. Korean also uses ideographs based on the set invented in China, called Hanja. Korean text requires over 6,000 Hanja characters. Hanja is used mostly to avoid confusion when Hangeul would be ambiguous. Hangeul characters are formed by combining consonants and vowels. After combining them, they can compose one syllable, which is a Hangeul character. Hangeul characters are often arranged in a square, so that the group takes up the same space as a Hanja character. Arabic numerals, Roman letters, and special symbol characters are also present in Korean text.

Thai Text

A Thai character can be defined as a column position on a display screen with four display cells. Each column position can have up to three characters. The composition of a display cell is based on the Thai character's classification. Some Thai characters can be composed with another character's classification. If they can be composed together, both characters are in the same cell. Otherwise, they are in separate cells.

Chinese Text

Chinese usually consists entirely of characters from the ideographic script called Hanzi. In the People's Republic of China (PRC) there are about 7000 commonly used Hanzi characters in GB2312 (zh locale) and more than 20,000 characters in the GBK (zh.GBK) locale. In Taiwan, current standards require more than 13000 characters; 6000 others have been recently standardized but are considered rare.

If a character is not a root character, it usually consists of two or more parts, two being most common. In two-part characters, one part generally represents meaning, and the other represents pronunciation. Occasionally both parts represent meaning. The radical is the most important element, and characters are traditionally arranged by radical, of which there are several hundred. A single sound can be represented by many different characters, which are not interchangeable in usage. A single character can have different sounds.

Some characters are more appropriate than others in a given context--the appropriate one is distinguished phonetically by the use of tones. By contrast, spoken Japanese and Korean lack tones.

Several phonetic systems represent Chinese. In the People's Republic of China the most common is pinyin, which uses roman characters and is widely employed in the West for place names such as Beijing. The Wade-Giles system is an older phonetic system, formerly used for place names such as Peking. In Taiwan zhuyin (or bopomofo), a phonetic alphabet with unique letter forms, is often used instead.

Commercial applications, particularly those that deal with people's names, need to consider the impact of codeset expansion. Many Chinese people have names containing characters that do not exist in any standard codeset. You need to provide space in unassigned codesets to deal with this issue.

- [Previous: Cultural Conventions](#)
- [Next: Keyboard Differences](#)

- © 2010, Oracle Corporation and/or its affiliates