Formats and Conventions Overview 2

Different countries in the world use completely different conventions for writing date, time, numbers, currency, delimiting words and phrases, and quoting material. You should not code these conventions directly into your application; an internationalized product is adapted to the appropriate formats by a localization package. Chapter 4, "Writing Internationalized Code" and Chapter 5, "Internationalizing OpenWindows" provide guidance on coding practices used to achieve these formats.

Formatting Differences

Time Formats

Table 2-1 shows some of the ways to write 11:59 p.m.

Table 2-1 International Time Formats

Locale	Format
Canadian	23:59
Finnish	23.59
German	23.59 Uhr
Norwegian	Kl 23.59
U.K.	11.59 PM

Time is represented by both a 12-hour clock and a 24-hour clock—sometimes known as "railroad time." The hour and minute separator can be either a colon (:) or a period (.). Some countries attach letters to the time indicating that this is a time, but these are not strictly necessary.

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, Newfoundland 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 2-2 shows some of the date formats used around the world. Of course, there is a good deal of variation even within countries, so these formats are not the final truth.

Table 2-2 International Date Formats

Locale	**Convention**	**Example**
Canadian (English)	yyyy-mm-dd	1989-08-13
Canadian (French)	yyyy-mm-dd	1989-08-13
Danish	dd/mm/yy	13/08/89
Finnish	dd.mm.yyyy	13.08.1989
French	dd/mm/yy	13/08/89
German	dd.mm.yy	13.08.89
Italian	dd.mm.yy	13.08.89
Norwegian	dd.mm.yy	13.08.89
Spanish	dd-mm-yy	13-08-89
Swedish	yyyy-mm-dd	1989-08-13
UK-English	dd/mm/yy	13/08/89
US-English	mm-dd-yy	08-13-89

Numbers

Decimal and Thousands Separators

The United Kingdom 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 thousands groups, many other countries use a period for this instead, and some countries separate thousands groups with a thin space. Table 2-3 shows some commonly used numeric formats.

Table 2-3 International Numeric Conventions

Locale	**Large Number**
Canadian (French)	4 294 967 295,00
Canadian (English)	4 294 967 295,00
Danish	4.294.967.295,00
Finnish	4.294.967.295,00
French	4.294.967.295,00
German	4 294 967 295,00
Italian	4.294.967.295,00
Norwegian	4.294.967.295,00
Spanish	4.294.967.295,00
Swedish	4.294.967.295,00
UK-English	4,294,967,295.00
US-English	4,294,967,295.00

Data files containing locale-specific formats will be misinterpreted when transferred to a system in a different locale. For example, a file containing numbers in a French format will not be 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 the UK and the US, but often spaces and semicolons are used. Certainly, international software never "hardwires" any number delimiter.

Currency

Currency units and presentation order vary greatly around the world. Table 2-4 shows monetary formats in some countries.

Table 2-4 International Monetary Conventions

Locale	**Currency**	**Example**
Canadian (English)	Dollar ($)	$1 234.56
Canadian (French)	Dollar ($)	1 234.56$
Danish	Kroner (kr)	kr.1.234,56
Finnish	Markka (mk)	1.234 mk
French	Franc (F)	F1.234,56
German	Deutsche Mark (DM)	1,234.56DM
Italian	Lira (L)	L1.234,56
Japanese	Yen (¥)	¥1,234
Norwegian	Krone (kr)	kr 1.234,56
Spanish	Peseta (Pts)	1.234,56Pts
Swedish	Krona (Kr)	1234.56KR
UK-English	Pound (£)	£1,234.56
US-English	Dollar ($)	$1,234.56

Note that 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 francs.

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

Word and Letter Differences

Word Delimiters

Usually, words are separated by a space character. In Japanese and Thai, however, there is often no delimiter between words.

Word Order

The order of words in phrases and sentences varies between languages. For instance, the order of the words “cat” and “black” in “a black cat” is reversed in the equivalent Spanish phrase, “uno gato negro.” And in French, the negatives “ne” and “pas” surround the word they negate, as in the phrase “I do not speak,” which in French is “Je ne parle pas.”

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.

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.

There are tens of thousands of Kanji 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 merely 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 avoid the use of Kanji completely, most Japanese readers find text containing Kanji easier to understand.

Korean Text

Korean is similar to Japanese in that Chinese-based ideograms, called Hanja, are mixed together with a phonetic alphabet, Hangul. Hanja is used mostly to avoid confusion when Hangul would be ambiguous.

Hangul characters are formed by combining 10 basic vowels and 14 consonants, 2 to 5 of which compose one syllable. Hangul characters are often arranged in a square like the four on a pair of dice, so that the group takes up the same space as a Hanja character.

Korean text requires over 6000 Hanja characters, plus about 96 Hangul characters.

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, although emerging standards number Hanzi in the tens of thousands. In the Republic of China (ROC or Taiwan) current standards require more than 13,000 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. The same sound can be represented by many different characters, which are not interchangeable in usage.

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.

There are several phonetic systems for representing Chinese. In mainland 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), an extensive 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 code set expansion. Many people in the ROC have names containing characters that do not exist in any standard code set. Space needs to be provided in unassigned code sets to deal with this issue.

Code Sets for x86

The default code set on the Solaris system for x86 is ISO-8859-1. IBM® DOS 437 code set is provided as an option in text mode; however, it is provided only at internationalization level 1. That is, if you choose to download IBM DOS 437 code set by typing:

```
loadfont -c 437
pcmapkeys -f /usr/share/lib/keyboards/437/en_US
```

there will be no support for nonstandard U.S. date, time, currency, numbers, units, and collation. There will be no support for non-English message and text presentation, and no multibyte character support. Therefore, non-MS-Windows™ users should use IBM DOS 437 code set only in the default C locale.

- You must be in the text mode to download the IBM code set, not the graphics mode.
- If you are not using the standard U.S. PC keyboard, replace `en_US` with the keyboard map related to your keyboard.
- To download the default code set in text mode, type:

```
loadfont -c 8859
pcmapkeys -f /usr/share/lib/keyboards/8859/en_US
```

- See the `loadfont` (1) and `pcmapkeys` (1) manual pages.

Keyboard Differences

Not all characters on the US keyboard appear on other keyboards. Similarly, other keyboards often contain many characters not visible on the US keyboard. However, the Compose key can be used to produce any character in the ISO Latin-1 code set on any keyboard that supports it. See Appendix B, "Compose Key Sequences" for a list of different keyboard layouts.

Other Differences

Punctuation

Both the position and the type of punctuation symbols can vary between languages. In Spanish, "¿" and "¡" appear at the beginnings of sentences, while in Finnish colons (:) can occur inside words.

Symbols

Commonly used symbols in one culture often have no meaning in another culture. Because the common U.S. rural mailbox, for example, does not exist in other countries, it would not make a universal MailTool icon.

Measurements

While most countries now use the metric system of measurement, the United States, parts of Canada, and the United Kingdom (albeit unofficially) still use the imperial system. The symbols for feet (‘) and inches (“) are not understood in all countries.

Gender

The spelling of adjectives, articles, and nouns are gender-dependent in some languages. In French, for example, “un petit gamin” and “une petite gamine” both mean “a cute kid.” The first expression, however, refers to a boy, and the second expression, to a girl. Also, neuter objects in English (“a computer” for example) have gender in other languages (“un ordinateur” is a masculine noun in French).

Titles and Addresses

Mr., Miss, Mrs., and Ms. are common titles in the US but are not used in many other countries.

Address formats differ from country to country. In many countries, the postal code includes letters as well as numbers.

Paper Sizes

Within each country a small number of paper sizes are commonly used, normally with one of those sizes being much more common than the others. Most countries follow ISO Standard 216 “Writing paper and certain classes of printed matter—Trimmed sizes—A and B series.”

Internationalized applications should not make assumptions about the page sizes available to them. The Solaris system provides no support for tracking output page size; this is the responsibility of the application program itself.

Table 2-5 Common International Page Sizes

Paper Type	Dimensions	Countries
ISO A4	21.0 cm by 29.7 cm	Everywhere except US
ISO A5	14.8 cm by 21.0 cm	Everywhere except US
JIS B4	25.9 cm by 36.65 cm	Japan
JIS B5	18.36 cm by 25.9 cm	Japan
US Letter	8.5 inch by 11 inches	US and Canada
US Legal	8.5 inch by 14 inches	US and Canada

Standard paper trays distributed with LaserWriter® and LaserWriter II printers support U.S. letter, U.S. legal, and A4 paper sizes. The SPARCprinter® paper tray supports all these sizes, in addition to B5.

Summary

International code does not contain any implicit cultural assumptions. Instead, it uses standardized interfaces that make use of installed localization packages. Read Chapters 4 and 5 to learn what these standard interfaces are and how you can use them to create internationalized products.