

# The Unicode® Standard

## Version 15.0 – Core Specification

To learn about the latest version of the Unicode Standard, see <https://www.unicode.org/versions/latest/>.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and the publisher was aware of a trademark claim, the designations have been printed with initial capital letters or in all capitals.

Unicode and the Unicode Logo are registered trademarks of Unicode, Inc., in the United States and other countries.

The authors and publisher have taken care in the preparation of this specification, but make no expressed or implied warranty of any kind and assume no responsibility for errors or omissions. No liability is assumed for incidental or consequential damages in connection with or arising out of the use of the information or programs contained herein.

The *Unicode Character Database* and other files are provided as-is by Unicode, Inc. No claims are made as to fitness for any particular purpose. No warranties of any kind are expressed or implied. The recipient agrees to determine applicability of information provided.

© 2022 Unicode, Inc.

All rights reserved. This publication is protected by copyright, and permission must be obtained from the publisher prior to any prohibited reproduction. For information regarding permissions, inquire at <https://www.unicode.org/reporting.html>. For information about the Unicode terms of use, please see <https://www.unicode.org/copyright.html>.

The Unicode Standard / the Unicode Consortium; edited by the Unicode Consortium. — Version 15.0.

Includes index.

ISBN 978-1-936213-32-0 (<https://www.unicode.org/versions/Unicode15.0.0/>)

1. Unicode (Computer character set) I. Unicode Consortium.

QA268.U545 2022

ISBN 978-1-936213-32-0

Published in Mountain View, CA

September 2022

## Chapter 22

# Symbols

The universe of symbols is rich and open-ended. The collection of encoded symbols in the Unicode Standard encompasses the following:

*Currency symbols*

*Letterlike symbols*

*Mathematical alphabets*

*Numerals*

*Superscript and subscript symbols*

*Mathematical symbols*

*Invisible mathematical operators*

*Technical symbols*

*Geometrical symbols*

*Miscellaneous symbols and dingbats*

*Pictographic symbols*

*Emoticons*

*Enclosed and square symbols*

Pictorial or graphic items for which there is no demonstrated need or strong desire to exchange in plain text are not encoded in the standard.

Combining marks may be used with symbols, particularly the set encoded at U+20D0.. U+20FF (see *Section 7.9, Combining Marks*).

Letterlike and currency symbols, as well as numerals, superscripts, and subscripts, are typically subject to the same font and style changes as the surrounding text. Where square and enclosed symbols occur in East Asian contexts, they generally follow the prevailing type styles.

Other symbols have an appearance that is independent of type style, or a more limited or altogether different range of type style variation than the regular text surrounding them. For example, mathematical alphanumeric symbols are typically used for mathematical variables; those letterlike symbols that are part of this set carry semantic information in their type style. This fact restricts—but does not completely eliminate—possible style variations. However, symbols such as mathematical operators can be used with any script or independent of any script.

Special invisible operator characters can be used to explicitly encode some mathematical operations, such as multiplication, which are normally implied by juxtaposition. This aids in automatic interpretation of mathematical notation.

In a bidirectional context (see Unicode Standard Annex #9, “Unicode Bidirectional Algorithm”), most symbol characters have no inherent directionality but resolve their directionality for display according to the Unicode Bidirectional Algorithm. For some symbols, such as brackets and mathematical operators whose image is not bilaterally symmetric, the

mirror image is used when the character is part of the right-to-left text stream (see *Section 4.7, Bidi Mirrored*).

Dingbats and optical character recognition characters are different from all other characters in the standard, in that they are encoded based primarily on their precise appearance.

Many symbols encoded in the Unicode Standard are intended to support legacy implementations and obsolescent practices, such as terminal emulation or other character mode user interfaces. Examples include box drawing components and control pictures.

A number of symbols are also encoded for emoji (“picture character,” or pictograph). Added initially for compatibility with the emoji sets encoded by several Japanese cell phone carriers as extensions of the JIS X 0208 character set, these pictographs continue to grow in usage and coverage. These symbols are interchanged as plain text, and are encoded in the Unicode Standard to support interoperability and widespread usage on mobile devices.

Other symbols—many of which are also pictographic—are encoded for compatibility with Webdings and Wingdings sets, or various e-mail systems, and to address other interchange requirements.

Many of the symbols encoded in Unicode can be used as operators or given some other syntactical function in a formal language syntax. For more information, see Unicode Standard Annex #31, “Unicode Identifier and Pattern Syntax.”

## 22.1 Currency Symbols

Currency symbols are intended to encode the customary symbolic signs used to indicate certain currencies in general text. These signs vary in shape and are often used for more than one currency. Not all currencies are represented by a special currency symbol; some use multiple-letter strings instead, such as “Sfr” for Swiss franc. Moreover, the abbreviations for currencies can vary by language. The Unicode Common Locale Data Repository (CLDR) provides further information; see *Appendix B.3, Other Unicode Online Resources*. Therefore, implementations that are concerned with the *exact* identity of a currency should not depend on an encoded currency sign character. Instead, they should follow standards such as the ISO 4217 three-letter currency codes, which are *specific* to currencies—for example, USD for U.S. dollar, CAD for Canadian dollar.

**Unification.** The Unicode Standard does not duplicate encodings where more than one currency is expressed with the same symbol. Many currency symbols are overstruck letters. There are therefore many minor variants, such as the U+0024 DOLLAR SIGN \$, with one or two vertical bars, or other graphical variation, as shown in *Figure 22-1*.

**Figure 22-1.** Alternative Glyphs for Dollar Sign



Claims that glyph variants of a certain currency symbol are used consistently to indicate a particular currency could not be substantiated upon further research. Therefore, the Unicode Standard considers these variants to be typographical and provides a single encoding for them. See ISO/IEC 10367, Annex B (informative), for an example of multiple renderings for U+00A3 POUND SIGN.

**Fonts.** Currency symbols are commonly designed to display at the same width as a digit (most often a European digit, U+0030..U+0039) to assist in alignment of monetary values in tabular displays. Like letters, they tend to follow the stylistic design features of particular fonts because they are used often and need to harmonize with body text. In particular, even though there may be more or less normative designs for the currency sign per se, as for the euro sign, type designers freely adapt such designs to make them fit the logic of the rest of their fonts. This partly explains why currency signs show more glyph variation than other types of symbols.

### **Currency Symbols: U+20A0–U+20CF**

This block contains currency symbols that are not encoded in other blocks. Contemporary and historic currency symbols encoded in other blocks are listed in *Table 22-1*. The table omits currency symbols known only from usage in ancient coinage, such as U+1017A GREEK TALENT SIGN and U+10196 ROMAN DENARIUS SIGN.

**Table 22-1.** Currency Symbols Encoded in Other Blocks

| Currency                      | Unicode Code Point |                                      |
|-------------------------------|--------------------|--------------------------------------|
| Dollar, milreis, escudo, peso | U+0024             | DOLLAR SIGN                          |
| Cent                          | U+00A2             | CENT SIGN                            |
| Pound and lira                | U+00A3             | POUND SIGN                           |
| General currency              | U+00A4             | CURRENCY SIGN                        |
| Yen or yuan                   | U+00A5             | YEN SIGN                             |
| Dutch florin                  | U+0192             | LATIN SMALL LETTER F WITH HOOK       |
| Dram                          | U+058F             | ARMENIAN DRAM SIGN                   |
| Afghani                       | U+060B             | AFGHANI SIGN                         |
| Rupee                         | U+09F2             | BENGALI RUPEE MARK                   |
| Rupee                         | U+09F3             | BENGALI RUPEE SIGN                   |
| Ana (historic)                | U+09F9             | BENGALI CURRENCY DENOMINATOR SIXTEEN |
| Ganda (historic)              | U+09FB             | BENGALI GANDA MARK                   |
| Rupee                         | U+0AF1             | GUJARATI RUPEE SIGN                  |
| Rupee                         | U+0BF9             | TAMIL RUPEE SIGN                     |
| Baht                          | U+0E3F             | THAI CURRENCY SYMBOL BAHT            |
| Riel                          | U+17DB             | KHMER CURRENCY SYMBOL RIEL           |
| German mark (historic)        | U+2133             | SCRIPT CAPITAL M                     |
| Yuan, yen, won, HKD           | U+5143             | CJK UNIFIED IDEOGRAPH-5143           |
| Yen                           | U+5186             | CJK UNIFIED IDEOGRAPH-5186           |
| Yuan                          | U+5706             | CJK UNIFIED IDEOGRAPH-5706           |
| Yuan, yen, won, HKD, NTD      | U+5713             | CJK UNIFIED IDEOGRAPH-5713           |
| Rupee                         | U+A838             | NORTH INDIC RUPEE MARK               |
| Rial                          | U+FD9C             | RIAL SIGN                            |

**Lira Sign.** A separate currency sign U+20A4 LIRA SIGN is encoded for compatibility with the HP Roman-8 character set, which is still widely implemented in printers. In general, U+00A3 POUND SIGN may be used for both the various currencies known as pound (or punt) and the currencies known as lira. Examples include the British pound sterling, the historic Irish punt, and the former lira currency of Italy. Until 2012, the lira sign was also used for the Turkish lira, but for current Turkish usage, see U+20BA TURKISH LIRA SIGN. As in the case of the dollar sign, the glyphic distinction between single- and double-bar versions of the sign is not indicative of a systematic difference in the currency.

**Dollar and Peso.** The dollar sign (U+0024) is used for many currencies in Latin America and elsewhere. In particular, this use includes current and discontinued Latin American peso currencies, such as the Mexican, Chilean, Colombian and Dominican pesos. However, the Philippine peso uses a different symbol found at U+20B1.

**Yen and Yuan.** Like the dollar sign and the pound sign, U+00A5 YEN SIGN has been used as the currency sign for more than one currency. The double-crossbar glyph is the official form for both the yen currency of Japan (JPY) and for the yuan (renminbi) currency of China (CNY). This is the case, despite the fact that some glyph standards historically specified a single-crossbar form, notably the OCR-A standard ISO 1073-1:1976, which influenced the representative glyph in various character set standards from China. In the Unicode Standard, U+00A5 YEN SIGN is intended to be the character for the currency sign for both the yen and the yuan, independent of the details of glyphic presentation.

As listed in *Table 22-1*, there are also a number of CJK ideographs to represent the words *yen* (or *en*) and *yuan*, as well as the Korean word *won*, and these also tend to overlap in use as currency symbols.

**Euro Sign.** The single currency for member countries of the European Economic and Monetary Union is the euro (EUR). The euro character is encoded in the Unicode Standard as U+20AC EURO SIGN.

**Indian Rupee Sign.** U+20B9 ₹ INDIAN RUPEE SIGN is the character encoded to represent the Indian rupee currency symbol introduced by the Government of India in 2010 as the official currency symbol for the Indian rupee (INR). It is distinguished from U+20A8 RUPEE SIGN, which is an older symbol not formally tied to any particular currency. There are also a number of script-specific rupee symbols encoded for historic usage by various scripts of India. See *Table 22-1* for a listing.

Rupee is also the common name for a number of currencies for other countries of South Asia and of Indonesia, as well as several historic currencies. It is often abbreviated using Latin letters, or may be spelled out or abbreviated in the Arabic script, depending on local conventions.

**Turkish Lira Sign.** The Turkish lira sign, encoded as U+20BA ₺ TURKISH LIRA SIGN, is a symbol representing the lira currency of Turkey. Prior to the introduction of the new symbol in 2012, the currency was typically abbreviated with the letters “TL”. The new symbol was selected by the Central Bank of Turkey from entries in a public contest and is quickly gaining common use, but the old abbreviation is also still in use.

**Ruble Sign.** The ruble sign, encoded as U+20BD ₽ RUBLE SIGN, was adopted as the official symbol for the currency of Russian Federation in 2013. Ruble is also used as the name of various currencies in Eastern Europe. In English, both spellings “ruble” and “rouble” are used.

**Lari Sign.** The lari sign, encoded as U+20BE ₾ LARI SIGN, was adopted as the official symbol for the currency of Georgia in 2014. The name *lari* is an old Georgian word denoting a hoard or property. The image for the lari sign is based on the letter U+10DA Ⴀ GEORGIAN LETTER LAS. The lari currency was established on October 2, 1995.

**Bitcoin Sign.** U+20BF ₿ BITCOIN SIGN represents the *bitcoin*, a cryptocurrency and payment system invented by programmers. A cryptocurrency such as the bitcoin works as a medium of exchange that uses cryptography to secure transactions and to control the creation of additional units of currency. It is categorized as a decentralized virtual or digital currency.

**Som Sign.** U+20C0 ₮ SOM SIGN was adopted as the official currency symbol of the Kyrgyz Republic on February 8, 2017. The som currency was introduced with bank notes on May 10, 1993 to replace the Soviet ruble. Coins were added later in 2008.

**Other Currency Symbols.** Additional forms of currency symbols are found in the Small Form Variants (U+FE50..U+FE6F) and the Halfwidth and Fullwidth Forms (U+FF00..U+FFEF) blocks. Those symbols have the General\_Category property value Currency\_Symbol (gc = Sc).

Ancient Greek and Roman monetary symbols, for such coins and values as the Greek *obol* or the Roman *denarius* and *as*, are encoded in the Ancient Greek Numbers (U+10140..U+1018F) and Ancient Symbols (U+10190..U+101CF) blocks. Those symbols denote values of weights and currencies, but are not used as regular currency symbols. As such, their General\_Category property value is Other\_Symbol (gc = So).

## 22.2 Letterlike Symbols

### **Letterlike Symbols: U+2100–U+214F**

Letterlike symbols are symbols derived in some way from ordinary letters of an alphabetic script. This block includes symbols based on Latin, Greek, and Hebrew letters. Stylistic variations of single letters are used for semantics in mathematical notation. See “Mathematical Alphanumeric Symbols” in this section for the use of letterlike symbols in mathematical formulas. Some letterforms have given rise to specialized symbols, such as U+211E PRESCRIPTION TAKE.

**Numero Sign.** U+2116 NUMERO SIGN is provided both for Cyrillic use, where it looks like №, and for compatibility with Asian standards, where it looks like №. Figure 22-2 illustrates a number of alternative glyphs for this sign. Instead of using a special symbol, French practice is to use an “N” or an “n”, according to context, followed by a superscript small letter “o” (N<sup>o</sup> or n<sup>o</sup>; plural N<sup>os</sup> or n<sup>os</sup>). Legacy data encoded in ISO/IEC 8859-1 (Latin-1) or other 8-bit character sets may also have represented the *numero sign* by a sequence of “N” followed by the *degree sign* (U+00B0 DEGREE SIGN). Implementations interworking with legacy data should be aware of such alternative representations for the *numero sign* when converting data.

**Figure 22-2.** Alternative Glyphs for Numero Sign

№ № N<sup>o</sup> № №

**Unit Symbols.** Several letterlike symbols are used to indicate units. In most cases, however, such as for SI units (Système International), the use of regular letters or other symbols is preferred. U+2113 SCRIPT SMALL L is commonly used as a non-SI symbol for the *liter*. Official SI usage prefers the regular *lowercase letter l*.

Three letterlike symbols have been given canonical equivalence to regular letters: U+2126 OHM SIGN, U+212A KELVIN SIGN, and U+212B ANGSTROM SIGN. In all three instances, the regular letter should be used. If text is normalized according to Unicode Standard Annex #15, “Unicode Normalization Forms,” these three characters will be replaced by their regular equivalents.

In normal use, it is better to represent degrees Celsius “°C” with a sequence of U+00B0 DEGREE SIGN + U+0043 LATIN CAPITAL LETTER C, rather than U+2103 DEGREE CELSIUS. For searching, treat these two sequences as identical. Similarly, the sequence U+00B0 DEGREE SIGN + U+0046 LATIN CAPITAL LETTER F is preferred over U+2109 DEGREE FAHRENHEIT, and those two sequences should be treated as identical for searching.

**Compatibility.** Some symbols are composites of several letters. Many of these composite symbols are encoded for compatibility with Asian and other legacy encodings. (See also “CJK Compatibility Ideographs” in Section 18.1, *Han*.) The use of these composite symbols



is discouraged where their presence is not required by compatibility. For example, in normal use, the symbols U+2121 TEL TELEPHONE SIGN and U+213B FAX FACSIMILE SIGN are simply spelled out.

In the context of East Asian typography, many letterlike symbols, and in particular composites, form part of a collection of compatibility symbols, the larger part of which is located in the CJK Compatibility block (see *Section 22.10, Enclosed and Square*). When used in this way, these symbols are rendered as “wide” characters occupying a full cell. They remain upright in vertical layout, contrary to the rotated rendering of their regular letter equivalents. See Unicode Standard Annex #11, “East Asian Width,” for more information.

Where the letterlike symbols have alphabetic equivalents, they collate in alphabetic sequence; otherwise, they should be treated as symbols. The letterlike symbols may have different directional properties than normal letters. For example, the four transfinite cardinal symbols (U+2135..U+2138) are used in ordinary mathematical text and do not share the strong right-to-left directionality of the Hebrew letters from which they are derived.

**Styles.** The letterlike symbols include some of the few instances in which the Unicode Standard encodes stylistic variants of letters as distinct characters. For example, there are instances of blackletter (*Fraktur*), double-struck, italic, and script styles for certain Latin letters used as mathematical symbols. The choice of these stylistic variants for encoding reflects their common use as distinct symbols. They form part of the larger set of mathematical alphanumeric symbols. For the complete set and more information on its use, see “Mathematical Alphanumeric Symbols” in this section. These symbols should not be used in ordinary, nonscientific texts.

Despite its name, U+2118 SCRIPT CAPITAL P is neither script nor capital—it is uniquely the Weierstrass elliptic function symbol derived from a calligraphic *lowercase* p. U+2113 SCRIPT SMALL L is derived from a special *italic* form of the *lowercase* letter *l* and, when it occurs in mathematical notation, is known as the symbol *ell*. Use U+1D4C1 MATHEMATICAL SCRIPT SMALL L as the *lowercase script l* for mathematical notation.

**Standards.** The Unicode Standard encodes letterlike symbols from many different national standards and corporate collections.

### ***Mathematical Alphanumeric Symbols: U+1D400–U+1D7FF***

The Mathematical Alphanumeric Symbols block contains a large extension of letterlike symbols used in mathematical notation, typically for variables. The characters in this block are intended for use only in mathematical or technical notation, and not in nontechnical text. When used with markup languages—for example, with Mathematical Markup Language (MathML)—the characters are expected to be used directly, instead of indirectly via entity references or by composing them from base letters and style markup.

**Words Used as Variables.** In some specialties, whole words are used as variables, not just single letters. For these cases, style markup is preferred because in ordinary mathematical notation the juxtaposition of variables generally implies multiplication, not word forma-

tion as in ordinary text. Markup not only provides the necessary scoping in these cases, but also allows the use of a more extended alphabet.

## Mathematical Alphabets

**Basic Set of Alphanumeric Characters.** Mathematical notation uses a basic set of mathematical alphanumeric characters, which consists of the following:

- The set of basic Latin digits (0–9) (U+0030..U+0039)
- The set of basic uppercase and lowercase Latin letters (a–z, A–Z)
- The uppercase Greek letters  $\text{A}–\Omega$  (U+0391..U+03A9), plus the nabla  $\nabla$  (U+2207) and the variant of theta  $\Theta$  given by U+03F4
- The lowercase Greek letters  $\alpha–\omega$  (U+03B1..U+03C9), plus the partial differential sign  $\partial$  (U+2202), and the six glyph variants  $\epsilon$ ,  $\vartheta$ ,  $\varepsilon$ ,  $\phi$ ,  $\varrho$ , and  $\varpi$ , given by U+03F5, U+03D1, U+03F0, U+03D5, U+03F1, and U+03D6, respectively

Only unaccented forms of the letters are used for mathematical notation, because general accents such as the acute accent would interfere with common mathematical diacritics. Examples of common mathematical diacritics that can interfere with general accents are the circumflex, macron, or the single or double dot above, the latter two of which are used in physics to denote derivatives with respect to the time variable. Mathematical symbols with diacritics are always represented by combining character sequences.

For some characters in the basic set of Greek characters, two variants of the same character are included. This is because they can appear in the same mathematical document with different meanings, even though they would have the same meaning in Greek text. (See “Variant Letterforms” in *Section 7.2, Greek*.)

**Additional Characters.** In addition to this basic set, mathematical notation uses the uppercase and lowercase digamma, in regular (U+03DC and U+03DD) and bold (U+1D7CA and U+1D7CB), and the four Hebrew-derived characters (U+2135..U+2138). Occasional uses of other alphabetic and numeric characters are known. Examples include U+0428 CYRILLIC CAPITAL LETTER SHA, U+306E HIRAGANA LETTER NO, and Eastern Arabic-Indic digits (U+06F0..U+06F9). However, these characters are used only in their basic forms, rather than in multiple mathematical styles.

**Dotless Characters.** In the Unicode Standard, the characters “i” and “j”, including their variations in the mathematical alphabets, have the `Soft_Dotted` property. Any conformant renderer will remove the dot when the character is followed by a nonspacing combining mark above. Therefore, using an individual mathematical italic *i* or *j* with math accents would result in the intended display. However, in mathematical equations an entire sub-expression can be placed underneath a math accent—for example, when a “wide hat” is placed on top of  $i+j$ , as shown in *Figure 22-3*.

In such a situation, a renderer can no longer rely simply on the presence of an adjacent combining character to substitute for the un-dotted glyph, and whether the dots should be

**Figure 22-3.** Wide Mathematical Accents

$$\widehat{i+j} = \hat{i} + \hat{j}$$

removed in such a situation is no longer predictable. Authors differ in whether they expect the dotted or dotless forms in that case.

In some documents *mathematical italic dotless i* or *j* is used explicitly without any combining marks, or even in contrast to the dotted versions. Therefore, the Unicode Standard provides the explicitly dotless characters U+1D6A4 MATHEMATICAL ITALIC SMALL DOTLESS I and U+1D6A5 MATHEMATICAL ITALIC SMALL DOTLESS J. These two characters map to the ISOAMSO entities *imath* and *jmath* or the T<sub>E</sub>X macros `\imath` and `\jmath`. These entities are, by default, always italic. The appearance of these two characters in the code charts is similar to the shapes of the entities documented in the ISO 9573-13 entity sets and used by T<sub>E</sub>X. The mathematical dotless characters do not have case mappings.

**Semantic Distinctions.** Mathematical notation requires a number of Latin and Greek alphabets that initially appear to be mere font variations of one another. The letter H can appear as plain or upright (H), bold (H), italic (H), as well as script, Fraktur, and other styles. However, in any given document, these characters have distinct, and usually unrelated, mathematical semantics. For example, a normal H represents a different variable from a bold H, and so on. If these attributes are dropped in plain text, the distinctions are lost and the meaning of the text is altered. Without the distinctions, the well-known Hamiltonian formula turns into the *integral* equation in the variable H as shown in *Figure 22-4*.

**Figure 22-4.** Style Variants and Semantic Distinctions in Mathematics

|                      |                                                     |
|----------------------|-----------------------------------------------------|
| Hamiltonian formula: | $\mathcal{H} = \int d\tau (\epsilon E^2 + \mu H^2)$ |
| Integral equation:   | $H = \int d\tau (\epsilon E^2 + \mu H^2)$           |

Mathematicians will object that a properly formatted integral equation requires all the letters in this example (except for the “d”) to be in italics. However, because the distinction between  $\mathcal{H}$  and  $H$  has been lost, they would recognize it as a fallback representation of an integral equation, and not as a fallback representation of the Hamiltonian. By encoding a separate set of alphabets, it is possible to preserve such distinctions in plain text.

**Mathematical Alphabets.** The sets of distinctly styled mathematical alphanumeric symbols are listed in *Table 22-2*.

The styles in *Table 22-2* represent those encountered in mathematical use. The plain letters have been unified with the existing characters in the Basic Latin and Greek blocks. There are 24 double-struck, italic, Fraktur, and script characters that already exist in the Letterlike Symbols block (U+2100..U+214F). These are explicitly unified with the characters in this block, and corresponding holes have been left in the mathematical alphabets.

**Table 22-2.** Mathematical Alphanumeric Symbols

| Math Style                 | Characters from Basic Set | Location    |
|----------------------------|---------------------------|-------------|
| plain (upright, serified)  | Latin, Greek, and digits  | BMP         |
| bold                       | Latin, Greek, and digits  | Plane 1     |
| italic                     | Latin and Greek           | Plane 1     |
| bold italic                | Latin and Greek           | Plane 1     |
| script (calligraphic)      | Latin                     | Plane 1/BMP |
| bold script (calligraphic) | Latin                     | Plane 1     |
| Fraktur                    | Latin                     | Plane 1     |
| bold Fraktur               | Latin                     | Plane 1     |
| double-struck              | Latin and digits          | Plane 1     |
| sans-serif                 | Latin and digits          | Plane 1     |
| sans-serif bold            | Latin, Greek, and digits  | Plane 1     |
| sans-serif italic          | Latin                     | Plane 1     |
| sans-serif bold italic     | Latin and Greek           | Plane 1     |
| monospace                  | Latin and digits          | Plane 1     |

The alphabets encoded in the Mathematical Alphanumeric Symbols block on Plane 1 represent the distinctions between different mathematically styled semantic alphabets, but the exact glyphs shown in the code charts are not intended to be prescriptive for actual mathematical font design. In particular, the script and double-struck styles show considerable variation across mathematical fonts.

Characters from the Mathematical Alphanumeric symbols block should *not* be used to represent styling of nonmathematical text.

**Script Style and Calligraphic Variants.** The mathematical script style, also referred to as the *calligraphic* style, has two widely recognized, specific variant styles: *chancery*, as exemplified by the glyph  $\mathcal{L}$ , and *roundhand*, as exemplified by the glyph  $\mathscr{L}$ . In most mathematical documents, the chancery calligraphic style and the roundhand calligraphic style are considered interchangeable. Accordingly, when the mathematical alphanumeric symbols were added to the Unicode Standard in Version 3.1, those two styles were unified. However, documentation subsequently emerged demonstrating that the regular (non-bold) uppercase Latin script characters occasionally show semantic contrasts between chancery style and roundhand style. To accommodate this usage, variation sequences have been defined, starting with Version 14.0, to distinguish chancery and roundhand variants. These variation sequences work as follows:

    An uppercase mathematical script style letter followed by U+FE00 displays in chancery style.

    An uppercase mathematical script style letter followed by U+FE01 displays in roundhand style.

Otherwise, an uppercase mathematical script style letter will display with the default for the font in use. The exact list of defined variation sequences can be found in the Standardized-

Variants.txt file in the Unicode Character Database. Note that variation sequences are not defined for the bold script alphabet, nor for lowercase letters of the regular script alphabet, as there is no evidence of systematic distinctive use of variant styles for those ranges.

The Unicode code charts use the roundhand calligraphic style to display mathematical script letters, including the various script symbols encoded in the Letterlike Symbols block on the BMP. That choice is less disruptive for legacy fonts, and is more consistent with the expected display for the occasional use of such letterlike symbols in nonmathematical contexts such as the use of U+2133 SCRIPT CAPITAL M for the pre-1949 symbol for the German currency unit *Mark*. By contrast, widely deployed specialty mathematical fonts such as Cambria Math and STIX Two Math default to the chancery calligraphic style, which is the specific script variant currently favored by mathematicians.

**Compatibility Decompositions.** All mathematical alphanumeric symbols have compatibility decompositions to the base Latin and Greek letters. This does not imply that the use of these characters is discouraged for mathematical use. Folding away such distinctions by applying the compatibility mappings is usually not desirable, as it loses the semantic distinctions for which these characters were encoded. See Unicode Standard Annex #15, “Unicode Normalization Forms.”

## Fonts Used for Mathematical Alphabets

Mathematicians place strict requirements on the *specific* fonts used to represent mathematical variables. Readers of a mathematical text need to be able to distinguish single-letter variables from each other, even when they do not appear in close proximity. They must be able to recognize the letter itself, whether it is part of the text or is a mathematical variable, and lastly which mathematical alphabet it is from.

**Fraktur.** The blackletter style is often referred to as *Fraktur* or *Gothic* in various sources. Technically, Fraktur and Gothic typefaces are distinct designs from blackletter, but any of several font styles similar in appearance to the forms shown in the charts can be used. In East Asian typography, the term *Gothic* is commonly used to indicate a sans-serif type style.

**Math Italics.** Mathematical variables are most commonly set in a form of italics, but not all italic fonts can be used successfully. For example, a math italic font should avoid a “tail” on the lowercase *italic letter z* because it clashes with subscripts. In common text fonts, the *italic letter v* and *Greek letter nu* are not very distinct. A rounded *italic letter v* is therefore preferred in a mathematical font. There are other characters that sometimes have similar shapes and require special attention to avoid ambiguity. Examples are shown in Figure 22-5.

**Hard-to-Distinguish Letters.** Not all sans-serif fonts allow an easy distinction between lowercase *l* and uppercase *I*, and not all monospaced (monowidth) fonts allow a distinction between the *letter l* and the *digit one*. Such fonts are not usable for mathematics. In Fraktur, the letters  $\mathfrak{l}$  and  $\mathfrak{1}$ , in particular, must be made distinguishable. Overburdened blackletter forms are inappropriate for mathematical notation. Similarly, the *digit zero* must be distinct from the *uppercase letter O* for all mathematical alphanumeric sets. Some characters

**Figure 22-5.** Easily Confused Shapes for Mathematical Glyphs

|                    |          |          |         |
|--------------------|----------|----------|---------|
| italic a           | <i>a</i> | <b>α</b> | alpha   |
| italic v (pointed) | <i>v</i> | <b>ν</b> | nu      |
| italic v (rounded) | <i>υ</i> | <b>υ</b> | upsilon |
| script X           | <i>ℵ</i> | <b>χ</b> | chi     |
| plain Y            | Y        | <b>Υ</b> | Upsilon |

are so similar that even mathematical fonts do not attempt to provide distinct glyphs for them. Their use is normally avoided in mathematical notation unless no confusion is possible in a given context—for example, *uppercase A* and *uppercase Alpha*.

**Font Support for Combining Diacritics.** Mathematical equations require that characters be combined with diacritics (dots, tilde, circumflex, or arrows above are common), as well as followed or preceded by superscripted or subscripted letters or numbers. This requirement leads to designs for *italic* styles that are less inclined and *script* styles that have smaller overhangs and less slant than equivalent styles commonly used for text such as wedding invitations.

**Double-Struck Characters.** The double-struck glyphs shown in earlier editions of the standard attempted to match the design used for all the other Latin characters in the standard, which is based on Times. The current set of fonts was prepared in consultation with the American Mathematical Society and leading mathematical publishers; it shows much simpler forms that are derived from the forms written on a blackboard. However, both serified and non-serified forms can be used in mathematical texts, and inline fonts are found in works published by certain publishers.

**Arabic Mathematical Alphabetic Symbols: U+1EE00–U+1EEFF**

The Arabic Mathematical Alphabetic Symbols block contains a set of characters used to write Arabic mathematical expressions. These symbols derive from a version of the Arabic alphabet which was widely used for many centuries and in a variety of contexts, such as in manuscripts and traditional print editions. The characters in this block follow the older, generic Semitic order (a, b, j, d...), differing from the order typically found in dictionaries (a, b, t, th...). These symbols are used by Arabic alphabet-based scripts, such as Arabic and Persian, and appear in the majority of mathematical handbooks published in the Middle East, Libya, and Algeria today.

In Arabic mathematical notation, much as in Latin-based mathematical text, style variation plays an important semantic role and must be retained in plain text. Hence Arabic styles for these mathematical symbols, which include tailed, stretched, looped, or double-struck forms, are encoded separately, and should not be handled at the font level. These mathematically styled symbols, which also include some isolated and initial-form Arabic letters,

are to be distinguished from the Arabic compatibility characters encoded in the Arabic Presentation Forms-B block.

**Shaping.** The Arabic Mathematical Symbols are not subject to shaping, unlike the Arabic letters in the Arabic block (U+0600..U+06FF).

**Large Operators.** Two operators are separately encoded: U+1EEF0 ARABIC MATHEMATICAL OPERATOR MEEM WITH HAH WITH TATWEEL, which denotes summation in Arabic mathematics, and U+1EEF1 ARABIC MATHEMATICAL OPERATOR HAH WITH DAL, which denotes limits in Persian mathematics. The glyphs for both of these characters stretch, based on the width of the text above or below them.

**Properties.** The characters in this block, although used as mathematical symbols, have the General\_Category value Lo. This property assignment for these letterlike symbols reflects the similar treatment for the alphanumeric mathematical symbols based on Latin and Greek letterforms.

## 22.3 Numerals

Many characters in the Unicode Standard are used to represent numbers or numeric expressions. Some characters are used exclusively in a numeric context; other characters can be used both as letters and numerically, depending on context. The notational systems for numbers are equally varied. They range from the familiar decimal notation to non-decimal systems, such as Roman numerals.

**Encoding Principles.** The Unicode Standard encodes sets of digit characters (or non-digit characters, as appropriate) for each script which has significantly distinct forms for numerals. As in the case of encoding of letters (and other units) for writing systems, the emphasis is on encoding the units of the written forms for numeric systems.

Sets of digits which differ by mathematical style are separately encoded, for use in mathematics. Such mathematically styled digits may carry distinct semantics which is maintained as a plain text distinction in the representation of mathematical expressions. This treatment of styled digits for mathematics parallels the treatment of styled alphabets for mathematics. See “Mathematical Alphabets” in *Section 22.2, Letterlike Symbols*.

Other font face distinctions for digits which do not have mathematical significance, such as the use of old style digits in running text, are not separately encoded. Other glyphic variations in digits and numeric characters are likewise not separately encoded. There are a few documented exceptions to this general rule. See “Glyph Variants of Decimal Digits” later in this section.

### *Decimal Digits*

A decimal digit is a digit that is used in decimal (radix 10) place value notation. The most widely used decimal digits are the European digits, encoded in the range from U+0030 DIGIT ZERO to U+0039 DIGIT NINE. Because of their early encoding history, these digits are also commonly known as *ASCII digits*. They are also known as *Western digits* or *Latin digits*. The European digits are used with a large variety of writing systems, including those whose own number systems are not decimal radix systems.

Many scripts also have their own decimal digits, which are separately encoded. Examples are the digits used with the Arabic script or those of the Indic scripts. *Table 22-3* lists scripts for which separate decimal digits are encoded, together with the section in the Unicode Standard which describes that script. The scripts marked with an asterisk (Arabic, Myanmar, and Tai Tham) have two or more sets of digits.

In the Unicode Standard, a character is formally classified as a decimal digit if it meets the conditions set out in “Decimal Digits” in *Section 4.6, Numeric Value* and has been assigned the property `Numeric_Type = Decimal`. The `Numeric_Type` property can be used to get the complete list of all decimal digits for any version of the Unicode Standard. (See `DerivedNumericType.txt` in the Unicode Character Database.)

When characters classified as decimal digits are used in sequences to represent decimal radix numerals, they are always stored most significant digit first. This convention includes



**Table 22-3.** Script-Specific Decimal Digits

| Script             | Section              |
|--------------------|----------------------|
| Adlam              | <i>Section 19.9</i>  |
| Ahom               | <i>Section 15.16</i> |
| Arabic*            | <i>Section 9.2</i>   |
| Balinese           | <i>Section 17.3</i>  |
| Bengali & Assamese | <i>Section 12.2</i>  |
| Bhaiksuki          | <i>Section 14.3</i>  |
| Brahmi             | <i>Section 14.1</i>  |
| Chakma             | <i>Section 13.11</i> |
| Cham               | <i>Section 16.10</i> |
| Devanagari         | <i>Section 12.1</i>  |
| Dives Akuru        | <i>Section 15.15</i> |
| Gujarati           | <i>Section 12.4</i>  |
| Gunjala Gondi      | <i>Section 13.15</i> |
| Gurmukhi           | <i>Section 12.3</i>  |
| Hanifi Rohingya    | <i>Section 16.14</i> |
| Javanese           | <i>Section 17.4</i>  |
| Kannada            | <i>Section 12.8</i>  |
| Kawi               | <i>Section 17.9</i>  |
| Kayah Li           | <i>Section 16.9</i>  |
| Khmer              | <i>Section 16.4</i>  |
| Khudawadi          | <i>Section 15.9</i>  |
| Lao                | <i>Section 16.2</i>  |
| Lepcha             | <i>Section 13.12</i> |
| Limbu              | <i>Section 13.6</i>  |
| Malayalam          | <i>Section 12.9</i>  |
| Masaram Gondi      | <i>Section 13.14</i> |
| Meetei Mayek       | <i>Section 13.7</i>  |
| Modi               | <i>Section 15.12</i> |

| Script                 | Section              |
|------------------------|----------------------|
| Mongolian              | <i>Section 13.5</i>  |
| Mro                    | <i>Section 13.8</i>  |
| Myanmar*               | <i>Section 16.3</i>  |
| Nag Mundari            | <i>Section 15.8</i>  |
| New Tai Lue            | <i>Section 16.6</i>  |
| Newa                   | <i>Section 13.3</i>  |
| N’Ko                   | <i>Section 19.4</i>  |
| Nyiakeng Puachue Hmong | <i>Section 16.12</i> |
| Ol Chiki               | <i>Section 13.10</i> |
| Oriya                  | <i>Section 12.5</i>  |
| Osmanya                | <i>Section 19.2</i>  |
| Pahawh Hmong           | <i>Section 16.11</i> |
| Saurashtra             | <i>Section 13.13</i> |
| Sharada                | <i>Section 15.3</i>  |
| Sinhala                | <i>Section 13.2</i>  |
| Sora Sompeng           | <i>Section 15.17</i> |
| Sundanese              | <i>Section 17.7</i>  |
| Tai Tham*              | <i>Section 16.7</i>  |
| Takri                  | <i>Section 15.4</i>  |
| Tamil                  | <i>Section 12.6</i>  |
| Tangsa                 | <i>Section 13.18</i> |
| Telugu                 | <i>Section 12.7</i>  |
| Thai                   | <i>Section 16.1</i>  |
| Tibetan                | <i>Section 13.4</i>  |
| Tirhuta                | <i>Section 15.11</i> |
| Vai                    | <i>Section 19.5</i>  |
| Wancho                 | <i>Section 13.16</i> |
| Warang Citi            | <i>Section 13.9</i>  |

decimal digits associated with scripts whose predominant layout direction is right-to-left. The visual layout of decimal radix numerals in bidirectional contexts depends on the interaction of their Bidi\_Class values with the Unicode Bidirectional Algorithm (UBA). In many cases, decimal digits share the same strong Bidi\_Class values with the letters of their script (“L” or “R”). A few common-use decimal digits, such as the ASCII digits and the Arabic script digits have special Bidi\_Class values that interact with dedicated rules for resolving the direction of numbers in the UBA. (See Unicode Standard Annex #9, “Unicode Bidirectional Algorithm.”)

The Unicode Standard does not specify which sets of decimal digits can or should be used with any particular writing system, language, or locale. However, the information provided in the Unicode Common Locale Data Repository (CLDR) contains information about which set or sets of digits are used with particular locales defined in CLDR. Numeral sys-

tems for a given locale require additional information, such as the appropriate decimal and grouping separators, the type of digit grouping used, and so on; that information is also supplied in CLDR.

**Exceptions.** There are several scripts with exceptional encodings for characters that are used as decimal digits. For the Arabic script, there are two sets of decimal digits encoded which have somewhat different glyphs and different directional properties. See “Arabic-Indic Digits” in *Section 9.2, Arabic* for a discussion of these two sets and their use in Arabic text. For the Myanmar script a second set of digits is encoded for the Shan language, and a third set of digits is encoded for the Tai Laing language. The Tai Tham script also has two sets of digits, which are used in different contexts.

**CJK Ideographs Used as Decimal Digits.** The CJK ideographs listed in *Table 4-5*, with numeric values in the range one through nine, can be used in decimal notations (with 0 represented by U+3007 IDEOGRAPHIC NUMBER ZERO). These ideographic digits are not coded in a contiguous sequence, nor do they occur in numeric order. Unlike other script-specific digits, they are not uniquely used as decimal digits. The same characters may be used in the traditional Chinese system for writing numbers, which is not a decimal radix system, but which instead uses numeric symbols for tens, hundreds, thousands, ten thousands, and so forth. See *Figure 22-6*, which illustrates two different ways the number 1,234 can be written with CJK ideographs.

**Figure 22-6.** CJK Ideographic Numbers

一千二百三十四  
or  
一 二 三 四

CJK numeric ideographs are also used in word compounds which are not interpreted as numbers. Parsing CJK ideographs as decimal numbers therefore requires information about the context of their use.

### **Other Digits**

**Hexadecimal Digits.** Conventionally, the letters “A” through “F”, or their lowercase equivalents are used with the ASCII decimal digits to form a set of hexadecimal digits. These characters have been assigned the `Hex_Digit` property. Although overlapping the letters and digits this way is not ideal from the point of view of numerical parsing, the practice is long standing; nothing would be gained by encoding a new, parallel, separate set of hexadecimal digits.

**Compatibility Digits.** There are several sets of compatibility digits in the Unicode Standard. *Table 22-4* provides a full list of compatibility digits.

Table 22-4. Compatibility Digits

| Description                                | Code Range(s)                      | Numeric Type     | Decomp Type | Section      |
|--------------------------------------------|------------------------------------|------------------|-------------|--------------|
| Fullwidth digits                           | FF10..FF19                         | Decimal          | Wide        | Section 18.5 |
| Bold digits                                | 1D7CE..1D7D7                       | Decimal          | Font        | Section 22.2 |
| Double struck                              | 1D7D8..1D7E1                       | Decimal          | Font        | Section 22.2 |
| Monospace digits                           | 1D7F6..1D7FF                       | Decimal          | Font        | Section 22.2 |
| Sans-serif digits                          | 1D7E2..1D7EB                       | Decimal          | Font        | Section 22.2 |
| Sans-serif bold digits                     | 1D7EC..1D7F5                       | Decimal          | Font        | Section 22.2 |
| Segmented digits                           | 1FBF0..1FBF9                       | Decimal          | Font        | Section 22.7 |
| Superscript digits                         | 2070, 00B9, 00B2, 00B3, 2074..2079 | Digit            | Super       | Section 22.4 |
| Subscript digits                           | 2080..2089                         | Digit            | Sub         | Section 22.4 |
| Circled digits                             | 24EA, 2080..2089                   | Digit            | Circle      |              |
| Parenthesized digits                       | 2474..247C                         | Digit            | Compat      |              |
| Digits plus full stop                      | 1F100, 2488..2490                  | Digit            | Compat      |              |
| Digits plus comma                          | 1F101..1F10A                       | Digit            | Compat      |              |
| Double circled digits                      | 24F5..24FD                         | Digit            | None        |              |
| Dingbat negative circled digits            | 2776..277E                         | Digit            | None        |              |
| Dingbat circled sans-serif digits          | 1F10B, 2780..2788                  | Numeric or Digit | None        |              |
| Dingbat negative circled sans-serif digits | 1F10C, 278A..2792                  | Numeric or Digit | None        |              |
| Segmented Digits                           | 1FBF0..1FBF9                       | Decimal          | Font        | Section 22.7 |

The fullwidth digits are simply wide presentation forms of ASCII digits, occurring in East Asian typographical contexts. They have compatibility decompositions to ASCII digits, have Numeric\_Type = Decimal, and should be processed as regular decimal digits.

The various mathematically styled digits in the range U+1D7CE..U+1D7F5 are specifically intended for mathematical use. They also have compatibility decompositions to ASCII digits and meet the criteria for Numeric\_Type = Decimal. Although they may have particular mathematical meanings attached to them, in most cases it would be safe for generic parsers to simply treat them as additional sets of decimal digits.

The segmented digits encoded in the range U+1FBF0..U+1FBF9 are used in legacy terminal applications, and are essentially just another styled set of ASCII digits. It is also safe for generic parsers to treat them as an additional set of decimal digits.

**Parsing of Superscript and Subscript Digits.** In the Unicode Character Database, superscript and subscript digits have not been given the General\_Category property value Decimal\_Number (gc = Nd); correspondingly, they have the Numeric\_Type property value Digit, rather than Decimal. This is to prevent superscripted expressions like 2<sup>3</sup> from being interpreted as 23 by simplistic parsers. More sophisticated numeric parsers, such as general mathematical expression parsers, should correctly identify these compatibility superscript and subscript characters as digits and interpret them appropriately. Note that the compatibility superscript digits are not encoded in a single, contiguous range.

For mathematical notation, the use of superscript or subscript styling of ASCII digits is preferred over the use of compatibility superscript or subscript digits. See Unicode Technical Report #25, “Unicode Support for Mathematics,” for more discussion of this topic.

**Numeric Bullets.** The other sets of compatibility digits listed in Table 22-4 are typically derived from East Asian legacy character sets, where their most common use is as numbered text bullets. Most occur as part of sets which extend beyond the value 9 up to 10, 12, or even 50. Most are also defective as sets of digits because they lack a value for 0. None is given the Numeric\_Type of Decimal. Only the basic set of simple circled digits is given compatibility decompositions to ASCII digits. The rest either have compatibility decompositions to digits plus punctuation marks or have no decompositions at all. Effectively, all of these numeric bullets should be treated as dingbat symbols with numbers printed on them; they should not be parsed as representations of numerals.

**Glyph Variants of Decimal Digits.** Some variations of decimal digits are considered glyph variants and are not separately encoded. These include the old style variants of digits, as shown in Figure 22-7. Glyph variants of the digit zero with a centered dot or a diagonal slash to distinguish it from the uppercase letter “O”, or of the digit seven with a horizontal bar to distinguish it from handwritten forms for the digit one, are likewise not separately encoded.

**Figure 22-7. Regular and Old Style Digits**

Regular Digits: 0 1 2 3 4 5 6 7 8 9

Old Style Digits: 0 1 2 3 4 5 6 7 8 9

In a few cases, such as for a small number of mathematical symbols, there may be a strong rationale for the unambiguous representation of a certain glyph variant of a decimal digit. In particular, the glyph variant of the digit zero with a short diagonal stroke “Ø” can be unambiguously represented with the standardized variation sequence <U+0030, U+FE00>.

Significant regional glyph variants for the Eastern-Arabic Digits U+06F0..U+06F9 also occur, but are not separately encoded. See Table 9-2 for illustrations of those variants.

**Accounting Numbers.** Accounting numbers are variant forms of digits or other numbers designed to deter fraud. They are used in accounting systems or on various financial instruments such as checks. These numbers often take shapes which cannot be confused with other digits or letters, and which are difficult to convert into another digit or number by adding on to the written form. When such numbers are clearly distinct characters, as opposed to merely glyph variants, they are separately encoded in the standard. The use of accounting numbers is particularly widespread in Chinese and Japanese, because the Han ideographs for one, two, and three have simple shapes that are easy to convert into other numbers by forgery. See Table 4-6, for a list of the most common alternate ideographs used as accounting numbers for the traditional Chinese numbering system.

Characters for accounting numbers are occasionally encoded separately for other scripts as well. For example, U+19DA NEW TAI LUE THAM DIGIT ONE is an accounting form for the digit one which cannot be confused with the vowel sign -aa and which cannot easily be converted into the digit for three.

## ***Non-Decimal Radix Systems***

A number of scripts have number systems that are not decimal place-value notations. Such systems are fairly common among traditional writing systems of South Asia. The following provides descriptions or references to descriptions of non-decimal radix systems elsewhere in the Standard.

***Ethiopic Numerals.*** The Ethiopic script contains digits and other numbers for a traditional number system which is not a decimal place-value notation. This traditional system does not use a zero. It is further described in *Section 19.1, Ethiopic*.

***Mende Kikakui Numerals.*** The Mende Kikakui script has a unique set of numerals, constituting a set of digits one through nine, used with a set of multiplier subscripts for powers of ten from 10 through 1,000,000. For more details on the structure of this numeral system, including examples, see *Section 19.8, Mende Kikakui*.

***Medefaidrin Numerals.*** The numerals used with the Medefaidrin script (see *Section 19.10, Medefaidrin*) constitute a novel, vigesimal radix system, with “digits” in the range 0 to 19. The Medefaidrin script is used only by a small community for religious purposes, so little is known about the practical use of these numerals.

***Mayan Numerals.*** Mayan writing used a set of vigesimal numerals, including a sign for zero. These signs are very well-known from Mayan calendrical inscriptions. They are striking in form, consisting of a series of horizontal bars with varying numbers of large dots above the bars, and so are easy to spot in inscriptions, amidst all the other hieroglyphic signs based on heads, animals, and so forth. The Mayan numerals are so well known, in fact, that they have gained a degree of modern re-use, appearing, for example, in page numbering of small documents published in Guatemala or the Yucatan. To accommodate such modern use of Mayan numerals, the full set has been encoded in the range U+1D2E0..U+1D2F3 in a dedicated Mayan Numerals block.

Until the analysis and encoding of the complex Mayan hieroglyphic script can be completed, these Mayan numerals stand by themselves. They are not given a Mayan Script property value, but are instead just treated as numeric symbols with the Script property Common.

***Kaktovik Numerals.*** Kaktovik numerals are vigesimal numerals including a sign for zero. These signs were devised by speakers of Iñupiaq in Kaktovik, Alaska for the counting systems of Inuit and Yupik languages. The top part of each numeral consists of up to three horizontal strokes, marking the fives, while the bottom part consists of up to four vertical strokes, marking the ones. The strokes are joined into a single continuous line.

Because the system is positional, for example U+1D2C1 KAKTOVIK NUMERAL ONE can indicate 1, 20, 400, 8,000, and so on, and U+1D2C5 KAKTOVIK NUMERAL FIVE indicates five times those amounts. Each Kaktovik numeral from 0 through 19 is encoded atomically.

**Cuneiform Numerals.** Sumero-Akkadian numerals were used for sexagesimal systems. There was no symbol for zero, but by Babylonian times, a place value system was in use. Thus the exact value of a digit depended on its position in a number. There was also ambiguity in numerical representation, because a symbol such as U+12079 CUNEIFORM SIGN DISH could represent either 1 or  $1 \times 60$  or  $1 \times (60 \times 60)$ , depending on the context. A numerical expression might also be interpreted as a sexagesimal fraction. So the sequence  $\langle 1, 10, 5 \rangle$  might be evaluated as  $1 \times 60 + 10 + 5 = 75$  or  $1 \times 60 \times 60 + 10 + 5 = 3615$  or  $1 + (10 + 5)/60 = 1.25$ . Many other complications arise in Cuneiform numeral systems, and they clearly require special processing distinct from that used for modern decimal radix systems. For more information, see *Section 11.1, Sumero-Akkadian*.

**Other Ancient Numeral Systems.** A number of other ancient numeral systems have characters encoded for them. Many of these ancient systems are variations on tallying systems. In numerous cases, the data regarding ancient systems and their use is incomplete, because of the fragmentary nature of the ancient text corpuses. Characters for numbers are encoded, however, to enable complete representation of the text which does exist.

Ancient Aegean numbers were used with the Linear A and Linear B scripts, as well as the Cypriot syllabary. They are described in *Section 8.2, Linear B*.

Many of the ancient Semitic scripts had very similar numeral systems which used tally-shaped numbers for one, two, and three, and which then grouped those, along with some signs for tens and hundreds, to form larger numbers. See the discussion of these systems in *Section 10.3, Phoenician* and, in particular, the discussion with examples of number formation in *Section 10.4, Imperial Aramaic*.

## ***Acrophonic Systems and Other Letter-based Numbers***

There are many instances of numeral systems, particularly historic ones, which use letters to stand for numbers. In some cases these systems may coexist with numeral systems using separate digits or other numbers. Two important sub-types are acrophonic systems, which assign numeric values based on the letters used for the initial sounds of number words, and alphabetic numerals, which assign numeric values based roughly on alphabetic order. A well-known example of a partially acrophonic system is the Roman numerals, which include c(entum) and m(ille) for 100 and 1000, respectively. The Greek Milesian numerals are an example of an alphabetic system, with alpha = 1, beta = 2, gamma = 3, and so forth.

In the Unicode Standard, although many letters in common scripts are known to be used for such letter-based numbers, they are not given numeric properties unless their *only* use is as an extension of an alphabet specifically for numbering. In most cases, the interpretation of letters or strings of letters as having numeric values is outside the scope of the standard.

**Roman Numerals.** For most purposes, it is preferable to compose the Roman numerals from sequences of the appropriate Latin letters. However, the uppercase and lowercase variants of the Roman numerals through 12, plus L, C, D, and M, have been encoded in the Number Forms block (U+2150..U+218F) for compatibility with East Asian standards. Unlike sequences of Latin letters, these symbols remain upright in vertical layout. Additionally, in certain locales, compact date formats use Roman numerals for the month, but may expect the use of a single character.

In identifiers, the use of Roman numeral symbols—particularly those based on a single letter of the Latin alphabet—can lead to spoofing. For more information, see Unicode Technical Report #36, “Unicode Security Considerations.”

U+2180 ROMAN NUMERAL ONE THOUSAND C D and U+216F ROMAN NUMERAL ONE THOUSAND can be considered to be glyphic variants of the same Roman numeral, but are distinguished because they are not generally interchangeable and because U+2180 cannot be considered to be a compatibility equivalent to the Latin letter M. U+2181 ROMAN NUMERAL FIVE THOUSAND and U+2182 ROMAN NUMERAL TEN THOUSAND are distinct characters used in Roman numerals; they do not have compatibility decompositions in the Unicode Standard. U+2183 ROMAN NUMERAL REVERSED ONE HUNDRED is a form used in combinations with C and/or I to form large numbers—some of which vary with single character number forms such as D, M, U+2181, or others. U+2183 is also used for the Claudian letter *antisigma*.

**Greek Numerals.** The ancient Greeks used a set of acrophonic numerals, also known as Attic numerals. These are represented in the Unicode Standard using capital Greek letters. A number of extensions for the Greek acrophonic numerals, which combine letterforms in odd ways, or which represent local regional variants, are separately encoded in the Ancient Greek Numbers block, U+10140..U+1018A.

Greek also has an alphabetic numeral system, called Milesian or Alexandrian numerals. These use the first third of the Greek alphabet to represent 1 through 9, the middle third for 10 through 90, and the last third for 100 through 900. U+0374 GREEK NUMERAL SIGN (the *dexia keraia*) marks letters as having numeric values in modern typography. U+0375 GREEK LOWER NUMERAL SIGN (the *aristeri keraia*) is placed on the left side of a letter to indicate a value in the thousands.

In Byzantine and other Greek manuscript traditions, numbers were often indicated by a horizontal line drawn above the letters being used as numbers. The Coptic script uses similar conventions. See *Section 7.3, Coptic*.

### ***Coptic Epact Numbers: U+102E0–U+102FF***

The Coptic epact numbers are elements of a decimal sign-value notation system used in some Coptic manuscripts. These numbers are referred to as “epact,” based on the Greek word ἐπακτός “imported.” They differ from the usual representation of numbers in Coptic texts, which consists of a system assigning numeric values directly to letters of the Coptic alphabet.

The Coptic epact numbers are considered to be historically derived from cursive forms of ordinary Coptic letters. They were developed in the 10th century CE by the Coptic community for administrative purposes. They are primarily attested in Coptic manuscripts written in Arabic, such as astronomical texts. They also appear in some accounting documents.

The numerical system for Coptic epact numbers is additive. The value of a numeric sequence consists of the sum of each number in the sequence. There is no character for zero. Instead, there are three sets of signs for the values 1 through 9, representing three orders: the digits, the tens, and the hundreds.

Numeric sequences are written from left to right, starting with the largest number at the left. For example, 25 is written **Ⲛⲉ** <U+102EB twenty, U+102E5 five>; 205 is written **Ⲫⲉ** <U+102F4 two hundred, U+102E5 five>; 250 is written **Ⲫⲟ** <U+102F4 two hundred, U+102EE fifty>. This order is followed even when Coptic epact numbers are embedded in right-to-left Arabic text.

Larger numbers are represented by applying a sublinear diacritical mark, U+102E0 COPTIC EPACT THOUSANDS MARK. Essentially, this mark multiplies the value of its base character by one thousand. Thus, when applied to symbols from the digits order, it represents thousands; when applied to symbols from the tens order, it represents ten thousands, and so on. A second application of the sublinear diacritic multiplies the base value by another factor of one thousand.

Ordinary Coptic numbers are often distinguished from Coptic letters by marking them with a line above. (See *Section 7.3, Coptic*.) A visually similar convention is also seen for Coptic epact numbers, where an entire numeric sequence may be marked with a wavy line above. This mark is represented by U+0605 ARABIC NUMBER MARK ABOVE. As when used with Arabic digits, ARABIC NUMBER MARK ABOVE *precedes* the sequence of Coptic epact numbers in the underlying representation, and is rendered across the top of the entire sequence for display.

### ***Rumi Numeral Symbols: U+10E60–U+10E7E***

Rumi, also known today as Fasi, is an numeric system used from the 10th to 17th centuries CE in a wide area, spanning from Egypt, across the Maghreb, to al-Andalus on the Iberian Peninsula. The Rumi numerals originate from the Coptic or Greek-Coptic tradition, but are not a positionally-based numbering system.

The numbers appear in foliation, chapter, and quire notations in manuscripts of religious, scientific, accounting and mathematical works. They also were used on astronomical instruments.

There is considerable variety in the Rumi glyph shapes over time: the digit “nine,” for example, appears in a theta shape in the early period. The glyphs in the code charts derive from a copy of a manuscript by Ibn Al-Banna (1256–1321), with glyphs that are similar to those found in 16th century manuscripts from the Maghreb.



## ***Siyaq Numerical Notation Systems***

There are a number of regional traditions for numerical notation systems known as Siyaq, derived from the Arabic word *siyāq*, meaning “order.” These traditions consist of specialized subsets of the Arabic script, formerly used in accounting and for general recording of numbers. A notable feature of Siyaq traditions is the use of stylized monograms of the Arabic names for numbers, rather than the ordinary Arabic-Indic digits.

Siyaq numbers represent units of a decimal positional system. The systems are additive—that is, the numeric value of a complete Siyaq number sequence consists of the sum of all the characters. There is no character for zero; instead, zero is represented inherently in the distinct numbers for the various decimal orders. Typically, there are distinctive numbers for the primary units, tens, hundreds, thousands, and ten thousands. The hundred thousands, millions, and higher orders are represented using unit marks and numbers from the smaller orders.

Siyaq numbers are written from right to left in the regular manner of the Arabic script. This orientation differs from the Arabic-Indic digits, which are written from left to right. In a Siyaq sequence, the largest number occurs first, and smaller units follow, laid out in visual order toward the left. An exception occurs for compound numbers of the tens and primary units; these are written transposed, with a “prefixed” form of the primary unit placed before the larger number.

***Ottoman Siyaq.*** The Ottoman, or Turkish, Siyaq numbers are encoded in the Ottoman Siyaq Numbers block (U+1ED00..U+1ED4F). These are also known as *Siyakat* numbers. The system contains several alternate forms for numbers, which may be historical retentions. These alternate forms are encoded as distinct characters for the numbers two through ten and for a few other numbers of higher orders.

The Ottoman Siyaq system includes a specialized multiplier character, U+1ED2E OTTOMAN SIYAQ MARRATAN (from the Arabic word *marratan*, “multiplier”). The multiplier is used in combination with *one hundred* and *one thousand* for expressing the millions and larger orders.

Ottoman Siyaq also uses a number of fractions. These fractions may be written in sequence after the number, or may be rendered beneath the number. Because of their distinctive shapes, two of the fractions are encoded as separate numeric symbols: U+1ED3C OTTOMAN SIYAQ FRACTION ONE HALF and U+1ED3D OTTOMAN SIYAQ FRACTION ONE SIXTH.

In some Ottoman Siyaq sources, a baseline dot indicates the end of a numerical sequence, and is placed after the last number. The dot can be represented either by U+002E FULL STOP or U+06D4 ARABIC FULL STOP, depending on the desired shape of the numerical terminator.

***Indic Siyaq.*** The Indic Siyaq tradition is known in India and other parts of South Asia as *raqm* or *rakam*, from the Arabic word *raqm*, meaning “account.” Indic Siyaq is encoded in the Indic Siyaq Numbers block (U+1EC70..U+1ECBF). Like other Siyaq traditions, Indic Siyaq uses stylized monograms of the Arabic names for numbers, but the numbers for large decimal orders are derived from words of Indic languages. The period during which Siyaq

was introduced in India is difficult to determine. The system was in common use under the Mughals by the 17th century, and remained in use into the middle of the 20th century.

There are two major styles of Siyaq used in India: the northern style and the “Deccani” or southern style. In general, the number forms and notation system of the two are identical. Minor points of difference lie in the orthography for the thousands, ten thousands, and lakhs.

There are also some minor style variations in writing tens of *lakhs* (millions) and tens of *crores* (hundred millions). For example, when writing the number ten *lakh*, one style may use a looped form of ten, looking like U+1EC7A INDIC SIYAQ NUMBER TEN, but another may use a straight form, looking like U+1EC95 INDIC SIYAQ NUMBER TEN THOUSAND. Such differences in style should be considered orthographic differences. The visual form seen in the documents being represented should be used to represent Indic Siyaq text. Thus, in one style the number ten *lakh* (one million) would be represented as <U+1EC7A, U+1ECA0>, but in another style as <U+1EC95, U+1ECA0>. Processes that interpret Indic Siyaq numbers should be aware of this irregular use of tens of thousands (U+1EC95..U+1EC9D) for tens when they appear before *lakhs* and *crores*.

The Indic Siyaq numbers are generally used within an Arabic script environment and within Urdu and Persian linguistic contexts. They may also occur in multilingual environments alongside other scripts. Arabic-Indic digits occasionally occur within Siyaq sequences, particularly for the representation of small currency units.

## CJK Numerals

**CJK Ideographic Traditional Numerals.** The traditional Chinese system for writing numerals is not a decimal radix system. It is decimal-based, but uses a series of decimal counter symbols that function somewhat like tallies. So for example, the representation of the number 12,346 in the traditional system would be by a sequence of CJK ideographs with numeric values as follows: <one, ten-thousand, two, thousand, three, hundred, four, ten, six>. See Table 4-5 for a list of all the CJK ideographs for digits and decimal counters used in this system. The traditional system is still in widespread use, not only in China and other countries where Chinese is used, but also in countries whose writing adopted Chinese characters—most notably, in Japan. In both China and Japan the traditional system now coexists with very common use of the European digits.

**Chinese Counting-Rod Numerals.** Counting-rod numerals were used in pre-modern East Asian mathematical texts in conjunction with counting rods used to represent and manipulate numbers. The counting rods were a set of small sticks, several centimeters long that were arranged in patterns on a gridded counting board. Counting rods and the counting board provided a flexible system for mathematicians to manipulate numbers, allowing for considerable sophistication in mathematics.

The specifics of the patterns used to represent various numbers using counting rods varied, but there are two main constants: Two sets of numbers were used for alternate columns; one set was used for the ones, hundreds, and ten-thousands columns in the grid, while the

other set was used for the tens and thousands. The shapes used for the counting-rod numerals in the Unicode Standard follow conventions from the Song dynasty in China, when traditional Chinese mathematics had reached its peak. Fragmentary material from many early Han dynasty texts shows different orientation conventions for the numerals, with horizontal and vertical marks swapped for the digits and tens places.

Zero was indicated by a blank square on the counting board and was either avoided in written texts or was represented with U+3007 IDEOGRAPHIC NUMBER ZERO. (Historically, U+3007 IDEOGRAPHIC NUMBER ZERO originated as a dot; as time passed, it increased in size until it became the same size as an ideograph. The actual size of U+3007 IDEOGRAPHIC NUMBER ZERO in mathematical texts varies, but this variation should be considered a font difference.) Written texts could also take advantage of the alternating shapes for the numerals to avoid having to explicitly represent zero. Thus 6,708 can be distinguished from 678, because the former would be 𠫪 𠫹 𠫹, whereas the latter would be 𠫹 𠫪 𠫹.

Negative numbers were originally indicated on the counting board by using rods of a different color. In written texts, a diagonal slash from lower right to upper left is overlaid upon the rightmost digit. On occasion, the slash might not be actually overlaid. U+20E5 COMBINING REVERSE SOLIDUS OVERLAY should be used for this negative sign.

The predominant use of counting-rod numerals in texts was as part of diagrams of counting boards. They are, however, occasionally used in other contexts, and they may even occur within the body of modern texts.

**Suzhou-Style Numerals.** The Suzhou-style numerals are CJK ideographic number forms encoded in the CJK Symbols and Punctuation block in the ranges U+3021..U+3029 and U+3038..U+303A.

The Suzhou-style numerals are modified forms of CJK ideographic numerals that are used by shopkeepers in China to mark prices. They are also known as “commercial forms,” “shop units,” or “grass numbers.” They are encoded for compatibility with the CNS 11643-1992 and Big Five standards. The forms for ten, twenty, and thirty, encoded at U+3038..U+303A, are also encoded as CJK unified ideographs: U+5341, U+5344, and U+5345, respectively. (For twenty, see also U+5EFE and U+5EFF.)

These commercial forms of Chinese numerals should be distinguished from the use of other CJK unified ideographs as accounting numbers to deter fraud. See *Table 4-6 in Section 4.6, Numeric Value*, for a list of ideographs used as accounting numbers.

Why are the Suzhou numbers called Hangzhou numerals in the Unicode names? No one has been able to trace this back. Hangzhou is a district in China that is near the Suzhou district, but the name “Hangzhou” does not occur in other sources that discuss these number forms.

## Fractions

The Number Forms block (U+2150..U+218F) contains a series of vulgar fraction characters, encoded for compatibility with legacy character encoding standards. These characters

are intended to represent both of the common forms of vulgar fractions: forms with a right-slanted division slash, such as  $\frac{1}{4}$ , as shown in the code charts, and forms with a horizontal division line, such as  $\frac{1}{4}$ , which are considered to be alternative glyphs for the same fractions, as shown in *Figure 22-8*. A few other vulgar fraction characters are located in the Latin-1 block in the range U+00BC..U+00BE.

**Figure 22-8.** Alternate Forms of Vulgar Fractions

$$\frac{1}{4} \quad \frac{1}{4}$$

The unusual fraction character, U+2189 VULGAR FRACTION ZERO THIRDS, is in origin a baseball scoring symbol from the Japanese television standard, ARIB STD B24. For baseball scoring, this character and the related fractions, U+2153 VULGAR FRACTION ONE THIRD and U+2154 VULGAR FRACTION TWO THIRDS, use the glyph form with the slanted division slash, and do not use the alternate stacked glyph form.

The vulgar fraction characters are given compatibility decompositions using U+2044 “/” FRACTION SLASH. Use of the *fraction slash* is the more generic way to represent fractions in text; it can be used to construct fractional number forms that are not included in the collections of vulgar fraction characters. For more information on the *fraction slash*, see “Other Punctuation” in *Section 6.2, General Punctuation*.

### **Common Indic Number Forms: U+A830–U+A83F**

The Common Indic Number Forms block contains characters widely used in traditional representations of fractional values in numerous scripts of North India, Pakistan and in some areas of Nepal. They are also regularly used in several scripts of South India, including Kannada. The fraction signs were used to write currency, weight, measure, time, and other units. Their use in written documents is attested from at least the 16th century CE and in texts printed as late as 1970. They are occasionally still used in a limited capacity.

The North Indic fraction signs represent fraction values of a base-16 notation system. There are atomic symbols for  $\frac{1}{16}$ ,  $\frac{2}{16}$ ,  $\frac{3}{16}$  and for  $\frac{1}{4}$ ,  $\frac{2}{4}$ , and  $\frac{3}{4}$ . Intermediate values such as  $\frac{5}{16}$  are written additively by using two of the atomic symbols:  $\frac{5}{16} = \frac{1}{4} + \frac{1}{16}$ , and so on. Some regional variation is found in the exact shape of the fraction signs used. For example, in Kannada, the fraction signs in the U+A833..U+A835 range are displayed with horizontal bars, instead of bars slanting upward to the right.

The signs for the fractions  $\frac{1}{4}$ ,  $\frac{1}{2}$ , and  $\frac{3}{4}$  sometimes take different forms when they are written independently, without a currency or quantity mark. These independent forms were used more generally in Maharashtra and Gujarat, and they appear in materials written and printed in the Devanagari and Gujarati scripts. The independent fraction signs are represented by using middle dots to the left and right of the regular fraction signs.

U+A836 NORTH INDIC QUARTER MARK is used in some regional orthographies to explicitly indicate fraction signs for  $1/4$ ,  $1/2$ , and  $3/4$  in cases where sequences of other marks could be ambiguous in reading.

This block also contains several other symbols that are not strictly number forms. They are used in traditional representation of numeric amounts for currency, weights, and other measures in the North Indic orthographies which use the fraction signs. U+A837 NORTH INDIC PLACEHOLDER MARK is a symbol used in currency representations to indicate the absence of an intermediate value. U+A839 NORTH INDIC QUANTITY MARK is a unit mark for various weights and measures.

The North Indic fraction signs are related to fraction signs that have specific forms and are separately encoded in some North Indic scripts. See, for example, U+09F4 BENGALI CURRENCY NUMERATOR ONE. Similar forms are attested for the Oriya script.

## 22.4 Superscript and Subscript Symbols

In general, the Unicode Standard does not attempt to describe the positioning of a character above or below the baseline in typographical layout. Therefore, the preferred means to encode superscripted letters or digits, such as “1<sup>st</sup>” or “DC00<sub>16</sub>”, is by style or markup in rich text. However, in some instances superscript or subscript letters are used as part of the plain text content of specialized phonetic alphabets, such as the Uralic Phonetic Alphabet. These superscript and subscript letters are mostly from the Latin or Greek scripts. These characters are encoded in other character blocks, along with other modifier letters or phonetic letters. In addition, superscript digits are used to indicate tone in transliteration of many languages. The use of *superscript two* and *superscript three* is common legacy practice when referring to units of area and volume in general texts.

### ***Superscripts and Subscripts: U+2070–U+209F***

A certain number of additional superscript and subscript characters are needed for round-trip conversions to other standards and legacy code pages. Most such characters are encoded in this block and are considered compatibility characters.

***Parsing of Superscript and Subscript Digits.*** In the Unicode Character Database, superscript and subscript digits have not been given the General\_Category property value `Decimal_Number` (gc = Nd), so as to prevent expressions like 2<sup>3</sup> from being interpreted like 23 by simplistic parsers. This should not be construed as preventing more sophisticated numeric parsers, such as general mathematical expression parsers, from correctly identifying these compatibility superscript and subscript characters as digits and interpreting them appropriately. See also the discussion of digits in *Section 22.3, Numerals*.

***Standards.*** Many of the characters in the Superscripts and Subscripts block are from character sets registered in the ISO International Register of Coded Character Sets to be Used With Escape Sequences, under the registration standard ISO/IEC 2375, for use with ISO/IEC 2022. Two MARC 21 character sets used by libraries include the digits, plus signs, minus signs, and parentheses.

***Superscripts and Subscripts in Other Blocks.*** The superscript digits one, two, and three are coded in the Latin-1 Supplement block to provide code point compatibility with ISO/IEC 8859-1. For a discussion of U+00AA FEMININE ORDINAL INDICATOR and U+00BA MASCULINE ORDINAL INDICATOR, see “Letters of the Latin-1 Supplement” in *Section 7.1, Latin*. U+2120 SERVICE MARK and U+2122 TRADE MARK SIGN are commonly used symbols that are encoded in the Letterlike Symbols block (U+2100..U+214F); they consist of sequences of two superscripted letters each.

For phonetic usage, there are a small number of superscript letters located in the Spacing Modifier Letters block (U+02B0..U+02FF) and a large number of superscript and subscript letters in the Phonetic Extensions block (U+1D00..U+1D7F) and in the Phonetic Extensions Supplement block (U+1D80..U+1DBF). Those superscript and subscript letters function as modifier letters. The subset of those characters that are superscripted contain the words “modifier letter” in their names, instead of “superscript.” The two superscript

Latin letters in the Superscripts and Subscripts block, U+2071 SUPERSCRIPT LATIN SMALL LETTER I and U+207F SUPERSCRIPT LATIN SMALL LETTER N are considered part of that set of modifier letters; the difference in the naming conventions for them is an historical artifact, and is not intended to convey a functional distinction in the use of those characters in the Unicode Standard.

There are also a number of superscript or subscript symbols encoded in the Spacing Modifier Letters block (U+02B0..U+02FF). These symbols also often have the words “modifier letter” in their names, but are distinguished from most modifier letters by having the General\_Category property value Sk. Like most modifier letters, the usual function of these superscript or subscript symbols is to indicate particular modifications of sound values in phonetic transcriptional systems. Characters such as U+02C2 MODIFIER LETTER LEFT ARROWHEAD or U+02F1 MODIFIER LETTER LOW LEFT ARROWHEAD should not be used to represent normal mathematical relational symbols such as U+003C “<” LESS-THAN SIGN in superscripted or subscripted expressions.

Finally, a small set of superscripted CJK ideographs, used for the Japanese system of syntactic markup of Classical Chinese text for reading, is located in the Kanbun block (U+3190..U+319F).

## 22.5 Mathematical Symbols

The Unicode Standard provides a large set of standard mathematical characters to support publications of scientific, technical, and mathematical texts on and off the Web. In addition to the mathematical symbols and arrows contained in the blocks described in this section, mathematical operators are found in the Basic Latin (ASCII) and Latin-1 Supplement blocks. These include U+002B PLUS SIGN, U+00D7 MULTIPLICATION SIGN and U+00F7 DIVISION SIGN, as well as U+003C GREATER THAN, U+003D EQUALS SIGN and U+003E LESS THAN. The *factorial operator* is unified with U+0021 EXCLAMATION MARK.

A few of the symbols from the Miscellaneous Technical, Miscellaneous Symbols, and Dingbats blocks, as well as characters from General Punctuation, are also used in mathematical notation. For Latin and Greek letters in special font styles that are used as mathematical variables, such as U+210B *ℋ* SCRIPT CAPITAL H, as well as the Hebrew letter *alef* used as the first transfinite cardinal symbol encoded by U+2135 *ℵ* ALEF SYMBOL, see “Letterlike Symbols” and “Mathematical Alphanumeric Symbols” in *Section 22.2, Letterlike Symbols*.

The repertoire of mathematical symbols in Unicode enables the display of virtually all standard mathematical symbols. Nevertheless, no collection of mathematical symbols can ever be considered complete; mathematicians and other scientists are continually inventing new mathematical symbols. More symbols will be added as they become widely accepted in the scientific communities.

**Semantics.** The same mathematical symbol may have different meanings in different sub-disciplines or different contexts. The Unicode Standard encodes only a single character for a single symbolic form. For example, the “+” symbol normally denotes addition in a mathematical context, but it might refer to concatenation in a computer science context dealing with strings, indicate incrementation, or have any number of other functions in given contexts. It is up to the application to distinguish such meanings according to the appropriate context. For some common mathematical symbols there are also local variations in usage. For example, in addition to its long history of use as punctuation mark, U+00D7 DIVISION SIGN is also used in certain cases to indicate negative numbers in several European countries. Where information is available about the usage (or usages) of particular symbols, it has been indicated in the character annotations in the code charts.

**Mathematical Property.** The mathematical (*math*) property is an informative property of characters that are used as operators in mathematical formulas. The mathematical property may be useful in identifying characters commonly used in mathematical text and formulas. However, a number of these characters have multiple usages and may occur with nonmathematical semantics. For example, U+002D HYPHEN-MINUS may also be used as a hyphen—and not as a mathematical minus sign. Other characters, including some alphabetic, numeric, punctuation, spaces, arrows, and geometric shapes, are used in mathematical expressions as well, but are even more dependent on the context for their identification. A list of characters with the mathematical property is provided in the Unicode Character Database.



For a classification of mathematical characters by typographical behavior and mapping to ISO 9573-13 entity sets, see Unicode Technical Report #25, “Unicode Support for Mathematics.”

### ***Mathematical Operators: U+2200–U+22FF***

The Mathematical Operators block includes character encodings for operators, relations, geometric symbols, and a few other symbols with special usages confined largely to mathematical contexts.

**Standards.** Many national standards’ mathematical operators are covered by the characters encoded in this block. These standards include such special collections as ANSI Y10.20, ISO 6862, ISO 8879, and portions of the collection of the American Mathematical Society, as well as the original repertoire of T<sub>E</sub>X.

**Encoding Principles.** Mathematical operators often have more than one meaning. Therefore the encoding of this block is intentionally rather shape-based, with numerous instances in which several semantic values can be attributed to the same Unicode code point. For example, U+2205 ∅ EMPTY SET may denote the mathematical concept of *empty set* or the linguistic concept of *null morpheme* or *phonological “zero.”* Similarly, U+2218 ∘ RING OPERATOR may be the equivalent of *white small circle* or *composite function* or *apl jot*. The Unicode Standard does not attempt to distinguish all possible semantic values that may be applied to mathematical operators or relation symbols.

The Unicode Standard does include many characters that appear to be quite similar to one another, but that may well convey different meanings in a given context. Conversely, mathematical operators, and especially relation symbols, may appear in various standards, handbooks, and fonts with a large number of purely graphical variants. Where variants were recognizable as such from the sources, they were not encoded separately.

Sometimes, specific glyph forms are chosen by notational style or are needed for contrast with other notation in the same document. For example, the symbol U+2205 ∅ EMPTY SET can be found in its slashed zero-shaped glyph form “Ø” in documents typeset in T<sub>E</sub>X, using the command `\emptyset`, or in contexts where it is contrasted with the semantically distinct digit zero.

For this and certain other well-established glyph variants of mathematical symbols, standardized variation sequences were added to the Unicode Standard. Thus, for example, the standardized variation sequence <U+2205, U+FE00> can be used to represent the variant “Ø” of the empty set symbol. To avoid the misuse of that sequence for the glyph variant of the digit zero with a short diagonal stroke “0”, the standardized variation sequence <U+0030, U+FE00> was separately specified for that digit glyph variant.

For relation symbols, the choice of a vertical or forward-slanting stroke typically indicating negation often seems to be an aesthetic one, but either slant might appear in a given context. However, a back-slanted stroke almost always has a distinct meaning compared to the forward-slanted stroke. See *Section 23.4, Variation Selectors*, for more information on some particular variants.

**Unifications.** Mathematical operators such as *implies*  $\Rightarrow$  and *if and only if*  $\Leftrightarrow$  have been unified with the corresponding arrows (U+21D2 RIGHTWARDS DOUBLE ARROW and U+2194 LEFT RIGHT ARROW, respectively) in the Arrows block.

The operator U+2208 ELEMENT OF is occasionally rendered with a taller shape than shown in the code charts. Mathematical handbooks and standards consulted treat these characters as variants of the same glyph. U+220A SMALL ELEMENT OF is a distinctively small version of the *element of* that originates in mathematical pi fonts.

The operators U+226B MUCH GREATER-THAN and U+226A MUCH LESS-THAN are sometimes rendered in a nested shape. The nested shapes are encoded separately as U+2AA2 DOUBLE NESTED GREATER-THAN and U+2AA1 DOUBLE NESTED LESS-THAN.

A large class of unifications applies to variants of relation symbols involving negation. Variants involving vertical or slanted *negation slashes* and *negation slashes* of different lengths are not separately encoded. For example, U+2288 NEITHER A SUBSET OF NOR EQUAL TO is the archetype for several different glyph variants noted in various collections.

In two instances in this block, essentially stylistic variants are separately encoded: U+2265 GREATER-THAN OR EQUAL TO is distinguished from U+2267 GREATER-THAN OVER EQUAL TO; the same distinction applies to U+2264 LESS-THAN OR EQUAL TO and U+2266 LESS-THAN OVER EQUAL TO. Further instances of the encoding of such stylistic variants can be found in the supplemental blocks of mathematical operators. The primary reason for such duplication is for compatibility with existing standards.

**Disunifications.** A number of mathematical operators have been disunified from related or similar punctuation characters, as shown in *Table 22-5*.

**Table 22-5.** Mathematical Operators Disunified from Punctuation

| Punctuation                                | Mathematical Operator    |
|--------------------------------------------|--------------------------|
| 002D - HYPHEN-MINUS                        | 2212 – MINUS SIGN        |
| 003F / SOLIDUS or <i>slash</i>             | 2215 / DIVISION SLASH    |
| 005C \ REVERSE SOLIDUS or <i>backslash</i> | 2216 \ SET MINUS         |
| 002A * ASTERISK                            | 2217 * ASTERISK OPERATOR |
| 25E6 ° WHITE BULLET                        | 2218 ° RING OPERATOR     |
| 2022 • BULLET                              | 2219 • BULLET OPERATOR   |
| 007C   VERTICAL LINE                       | 2223   DIVIDES           |
| 2016    DOUBLE VERTICAL LINE               | 2225    PARALLEL TO      |
| 003A : COLON                               | 2236 : RATIO             |
| 007E ~ TILDE                               | 223C ~ TILDE OPERATOR    |
| 00B7 · MIDDLE DOT                          | 22C5 · DOT OPERATOR      |

These disunifications support specific mathematical semantics, as well as some significant display differences between the punctuation marks and the operators. Mathematical oper-

ators render on the math centerline, rather than the text baseline. Additionally, the angle or length of the operator counterparts of certain slashes or bars may differ from the corresponding punctuation marks. For certain pairs, such as COLON and RATIO, there are distinctions in the behavior of inter-character spacing; RATIO is rendered as a relational operator which takes visible space on both sides, whereas the punctuation mark COLON does not require such additional space in rendering.

The distinction between MIDDLE DOT and DOT OPERATOR deserves special consideration. DOT OPERATOR is preferred for mathematical use, where it signifies multiplication. This allows for rendering consistent with other mathematical operators, with unambiguous character properties and mathematical semantics. MIDDLE DOT is a legacy punctuation mark, with multiple uses, and with quite variable layout in different fonts. For the typographical convention of a *raised decimal point*, in contexts where simple layout is the priority and where automated parsing of decimal expressions is not required, MIDDLE DOT is the preferred representation.

In cases where there ordinarily is no rendering distinction between a punctuation mark and its use in mathematics, such as for U+0021 ! EXCLAMATION MARK used for *factorial* or for U+002E FULL STOP used for a *baseline decimal point*, there is no disunification, and only a single character has been encoded.

**Greek-Derived Symbols.** Several mathematical operators derived from Greek characters have been given separate encodings because they are used differently from the corresponding letters. These operators may occasionally occur in context with Greek-letter variables. They include U+2206  $\Delta$  INCREMENT, U+220F  $\prod$  N-ARY PRODUCT, and U+2211  $\sum$  N-ARY SUMMATION. The latter two are large operators that take limits.

Other duplicated Greek characters are those for U+00B5  $\mu$  MICRO SIGN in the Latin-1 Supplement block, U+2126  $\Omega$  OHM SIGN in Letterlike Symbols, and several characters among the APL functional symbols in the Miscellaneous Technical block. Most other Greek characters with special mathematical semantics are found in the Greek block because duplicates were not required for compatibility. Additional sets of mathematical-style Greek alphabets are found in the Mathematical Alphanumeric Symbols block.

**N-ary Operators.** N-ary operators are distinguished from binary operators by their larger size and by the fact that in mathematical layout, they take limit expressions.

**Invisible Operators.** In mathematics, some operators or punctuation are often implied but not displayed. For a set of invisible operators that can be used to mark these implied operators in the text, see Section 22.6, *Invisible Mathematical Operators*.

**Minus Sign.** U+2212 “-” MINUS SIGN is a mathematical operator, to be distinguished from the ASCII-derived U+002D “-” HYPHEN-MINUS, which may look the same as a minus sign or be shorter in length. (For a complete list of dashes in the Unicode Standard, see Table 6-3.) U+22EE..U+22F1 are a set of ellipses used in matrix notation. U+2052 “/” COMMERCIAL MINUS SIGN is a specialized form of the minus sign. Its use is described in Section 6.2, *General Punctuation*.

**Delimiters.** Many mathematical delimiters are unified with punctuation characters. See *Section 6.2, General Punctuation*, for more information. Some of the set of ornamental brackets in the range U+2768..U+2775 are also used as mathematical delimiters. See *Section 22.9, Miscellaneous Symbols*. See also *Section 22.7, Technical Symbols*, for specialized characters used for large vertical or horizontal delimiters.

**Bidirectional Layout.** In a bidirectional context, with the exception of arrows, the glyphs for mathematical operators and delimiters are adjusted as described in Unicode Standard Annex #9, “Unicode Bidirectional Algorithm.” See *Section 4.7, Bidi Mirrored*, and “Paired Punctuation” in *Section 6.2, General Punctuation*.

**Other Elements of Mathematical Notation.** In addition to the symbols in these blocks, mathematical and scientific notation makes frequent use of arrows, punctuation characters, letterlike symbols, geometrical shapes, and miscellaneous and technical symbols.

For an extensive discussion of mathematical alphanumeric symbols, see *Section 22.2, Letterlike Symbols*. For additional information on all the mathematical operators and other symbols, see Unicode Technical Report #25, “Unicode Support for Mathematics.”

## ***Supplements to Mathematical Symbols and Arrows***

The Unicode Standard defines a number of additional blocks to supplement the repertoire of mathematical operators and arrows. These additions are intended to extend the Unicode repertoire sufficiently to cover the needs of such applications as MathML, modern mathematical formula editing and presentation software, and symbolic algebra systems.

**Standards.** MathML, an XML application, is intended to support the full legacy collection of the ISO mathematical entity sets. Accordingly, the repertoire of mathematical symbols for the Unicode Standard has been supplemented by the full list of mathematical entity sets in ISO TR 9573-13, *Public entity sets for mathematics and science*. An additional repertoire was provided from the amalgamated collection of the STIX Project (Scientific and Technical Information Exchange). That collection includes, but is not limited to, symbols gleaned from mathematical publications by experts of the American Mathematical Society and symbol sets provided by Elsevier Publishing and by the American Physical Society.

## ***Supplemental Mathematical Operators: U+2A00–U+2AFF***

The Supplemental Mathematical Operators block contains many additional symbols to supplement the collection of mathematical operators.

## ***Miscellaneous Mathematical Symbols-A: U+27C0–U+27EF***

The Miscellaneous Mathematical Symbols-A block contains symbols that are used mostly as operators or delimiters in mathematical notation.

**Mathematical Brackets.** The mathematical white square brackets, angle brackets, double angle brackets, and tortoise shell brackets encoded at U+27E6..U+27ED are intended for ordinary mathematical use of these particular bracket types. They are unambiguously nar-

row, for use in mathematical and scientific notation, and should be distinguished from the corresponding wide forms of white square brackets, angle brackets, and double angle brackets used in CJK typography. (See the discussion of the CJK Symbols and Punctuation block in *Section 6.2, General Punctuation*.) Note especially that the “bra” and “ket” angle brackets (U+2329 LEFT-POINTING ANGLE BRACKET and U+232A RIGHT-POINTING ANGLE BRACKET, respectively) are deprecated. Their use is strongly discouraged, because of their canonical equivalence to CJK angle brackets. This canonical equivalence is likely to result in unintended spacing problems if these characters are used in mathematical formulae.

The flattened parentheses encoded at U+27EE..U+27EF are additional, specifically-styled mathematical parentheses. Unlike the mathematical and CJK brackets just discussed, the flattened parentheses do not have corresponding wide CJK versions which they would need to be contrasted with.

**Long Division.** U+27CC LONG DIVISION is an operator intended for the representation of long division expressions, as may be seen in elementary and secondary school mathematical textbooks, for example. In use and rendering it shares some characteristics with U+221A SQUARE ROOT; in rendering, the top bar may be stretched to extend over the top of the denominator of the division expression. Full support of such rendering may, however, require specialized mathematical software.

**Fractional Slash and Other Diagonals.** U+27CB MATHEMATICAL RISING DIAGONAL and U+27CD MATHEMATICAL FALLING DIAGONAL are limited-use mathematical symbols, to be distinguished from the more widely used solidi and reverse solidi operators encoded in the Basic Latin, Mathematical Operators, and Miscellaneous Mathematical Symbols-B blocks. Their glyphs are invariably drawn at a 45 degree angle, instead of the more upright slants typical for the solidi operators. The box drawing characters U+2571 and U+2572, whose glyphs may also be found at a 45 degree angle in some fonts, are not intended to be used as mathematical symbols. One usage recorded for U+27CB and U+27CD is in the notation for spaces of double cosets. The former corresponds to the LaTeX entity `\diagup` and the latter to `\diagdown`.

### **Miscellaneous Mathematical Symbols-B: U+2980–U+29FF**

The Miscellaneous Mathematical Symbols-B block contains miscellaneous symbols used for mathematical notation, including fences and other delimiters. Some of the symbols in this block may also be used as operators in some contexts.

**Wiggly Fence.** U+29D8 LEFT WIGGLY FENCE has a superficial similarity to U+FE34 PRESENTATION FORM FOR VERTICAL WAVY LOW LINE. The latter is a wiggly sidebar character, intended for legacy support as a style of underlining character in a vertical text layout context; it has a compatibility mapping to U+005F LOW LINE. This represents a very different usage from the standard use of fence characters in mathematical notation.

**Tiny and Miny.** U+29FE TINY and U+29FF MINY are unary mathematical operators used in combinatorial game theory. TINY yields an infinitesimal positive value, while MINY yields

an infinitesimal negative value. The glyphs for TINY and MINY resemble the plus sign and minus sign, respectively, but should be shown distinctly, with thickened ends to their bars.

### **Miscellaneous Symbols and Arrows: U+2B00–U+2B7F**

The Miscellaneous Symbols and Arrows block contains more mathematical symbols and arrows. The arrows in this block extend and complete sets of arrows in other blocks. The other mathematical symbols complement various sets of geometric shapes. For a discussion of the use of such shape symbols in mathematical contexts, see “Geometric Shapes: U+25A0–U+25FF” and “Geometric Shapes Extended: U+1F780–U+1F7FF” in Section 22.8, *Geometrical Symbols*.

This block also contains various types of generic symbols. These complement the set of symbols in the Miscellaneous Symbols block, U+2600..U+26FF.

### **Arrows: U+2190–U+21FF**

Arrows are used for a variety of purposes: to imply directional relation, to show logical derivation or implication, and to represent the cursor control keys.

Accordingly, the Unicode Standard includes a fairly extensive set of generic arrow shapes, especially those for which there are established usages with well-defined semantics. It does not attempt to encode every possible stylistic variant of arrows separately, especially where their use is mainly decorative. For most arrow variants, the Unicode Standard provides encodings in the two horizontal directions, often in the four cardinal directions. For the single and double arrows, the Unicode Standard provides encodings in eight directions.

**Bidirectional Layout.** In bidirectional layout, arrows are not automatically mirrored, because the direction of the arrow could be relative to the text direction or relative to an absolute direction. Therefore, if text is copied from a left-to-right to a right-to-left context, or vice versa, the character code for the desired arrow direction in the new context must be used. For example, it might be necessary to change U+21D2 RIGHTWARDS DOUBLE ARROW to U+21D0 LEFTWARDS DOUBLE ARROW to maintain the semantics of “implies” in a right-to-left context. For more information on bidirectional layout, see Unicode Standard Annex #9, “Unicode Bidirectional Algorithm.”

**Standards.** The Unicode Standard encodes arrows from many different international and national standards as well as corporate collections.

**Unifications.** Arrows expressing mathematical relations have been encoded in the Arrows block as well as in the supplemental arrows blocks. An example is U+21D2  $\Rightarrow$  RIGHTWARDS DOUBLE ARROW, which may be used to denote *implies*. Where available, such usage information is indicated in the annotations to individual characters in the code charts. However, because the arrows have such a wide variety of applications, there may be several semantic values for the same Unicode character value.

## Supplemental Arrows

The Supplemental Arrows-A (U+27F0..U+27FF), Supplemental Arrows-B (U+2900..U+297F), Miscellaneous Symbols and Arrows (U+2B00..U+2BFF), and Supplemental Arrows-C (U+1F800..U+1F8FF) blocks contain a large repertoire of arrows to supplement the main set in the Arrows block. Many of the supplemental arrows in the Miscellaneous Symbols and Arrows block, particularly in the range U+2B30..U+2B4C, are encoded to ensure the availability of left-right symmetric pairs of less common arrows, for use in bidirectional layout of mathematical text.

**Long Arrows.** The long arrows encoded in the range U+27F5..U+27FF map to standard SGML entity sets supported by MathML. Long arrows represent distinct semantics from their short counterparts, rather than mere stylistic glyph differences. For example, the shorter forms of arrows are often used in connection with limits, whereas the longer ones are associated with mappings. The use of the long arrows is so common that they were assigned entity names in the ISOAMSA entity set, one of the suite of mathematical symbol entity sets covered by the Unicode Standard.

## Standardized Variants of Mathematical Symbols

These mathematical variants are all produced with the addition of U+FE00 VARIATION SELECTOR-1 (VS1) to mathematical operator base characters. The valid combinations are listed in the file StandardizedVariants.txt in the Unicode Character Database. All combinations not listed there are unspecified and are reserved for future standardization; no conformant process may interpret them as standardized variants.

**Change in Representative Glyphs for U+2278 and U+2279.** In Version 3.2 of the Unicode Standard, the representative glyphs for U+2278 NEITHER LESS-THAN NOR GREATER-THAN and U+2279 NEITHER GREATER-THAN NOR LESS-THAN were changed from using a vertical cancellation to using a slanted cancellation. This change was made to match the long-standing canonical decompositions for these characters, which use U+0338 COMBINING LONG SOLIDUS OVERLAY. The symmetric forms using the vertical stroke continue to be acceptable glyph variants. Using U+2276 LESS-THAN OR GREATER-THAN or U+2277 GREATER-THAN OR LESS-THAN with U+20D2 COMBINING LONG VERTICAL LINE OVERLAY will display these variants explicitly. Unless fonts are created with the intention to add support for both forms, there is no need to revise the glyphs in existing fonts; the glyphic range implied by using the base character code alone encompasses both shapes. For more information, see *Section 23.4, Variation Selectors*.

## 22.6 Invisible Mathematical Operators

In mathematics, some operators and punctuation are often implied but not displayed. The General Punctuation block contains several special format control characters known as *invisible operators*, which can be used to make such operators explicit for use in machine interpretation of mathematical expressions. Use of invisible operators is optional and is intended for interchange with math-aware programs.

A more complete discussion of mathematical notation can be found in Unicode Technical Report #25, “Unicode Support for Mathematics.”

**Invisible Separator.** U+2063 INVISIBLE SEPARATOR (also known as *invisible comma*) is intended for use in index expressions and other mathematical notation where two adjacent variables form a list and are not implicitly multiplied. In mathematical notation, commas are not always explicitly present, but they need to be indicated for symbolic calculation software to help it disambiguate a sequence from a multiplication. For example, the double  $ij$  subscript in the variable  $a_{ij}$  means  $a_{i,j}$ —that is, the  $i$  and  $j$  are separate indices and not a single variable with the name  $ij$  or even the product of  $i$  and  $j$ . To represent the implied list separation in the subscript  $ij$ , one can insert a nondisplaying *invisible separator* between the  $i$  and the  $j$ . In addition, use of the invisible comma would hint to a math layout program that it should typeset a small space between the variables.

**Invisible Multiplication.** Similarly, an expression like  $mc^2$  implies that the mass  $m$  multiplies the square of the speed  $c$ . To represent the implied multiplication in  $mc^2$ , one inserts a nondisplaying U+2062 INVISIBLE TIMES between the  $m$  and the  $c$ . Another example can be seen in the expression  $f^{ij}(\cos(ab))$ , which has the same meaning as  $f^{ij}(\cos(a \times b))$ , where  $\times$  represents *multiplication*, not the *cross product*. Note that the spacing between characters may also depend on whether the adjacent variables are part of a list or are to be concatenated (that is, multiplied).

**Invisible Plus.** The invisible plus operator, U+2064 INVISIBLE PLUS, is used to unambiguously represent expressions like  $3\frac{1}{4}$  which occur frequently in school and engineering texts. To ensure that  $3\frac{1}{4}$  means 3 plus  $\frac{1}{4}$ —in uses where it is not possible to rely on a human reader to disambiguate the implied intent of juxtaposition—the invisible plus operator is used. In such uses, not having an operator at all would imply multiplication.

**Invisible Function Application.** U+2061 FUNCTION APPLICATION is used for an implied function dependence, as in  $f(x + y)$ . To indicate that this is the function  $f$  of the quantity  $x + y$  and not the expression  $fx + fy$ , one can insert the nondisplaying *function application symbol* between the  $f$  and the left parenthesis.



## 22.7 Technical Symbols

### ***Control Pictures: U+2400–U+243F***

The need to show the presence of the C0 control codes unequivocally when data are displayed has led to conventional representations for these nongraphic characters.

***Code Points for Pictures for Control Codes.*** By definition, control codes themselves are manifested only by their action. However, it is sometimes necessary to show the position of a control code within a data stream. Conventional illustrations for the ASCII C0 control codes have been developed—but the characters U+2400..U+241F and U+2424 are intended for use as unspecified graphics for the corresponding control codes. This choice allows a particular application to use *any* desired pictorial representation of the given control code. It assumes that the particular pictures used to represent control codes are often specific to different systems and are rarely the subject of text interchange between systems.

***Pictures for ASCII Space.*** By definition, the SPACE is a blank graphic. Conventions have also been established for the visible representation of the space. Three specific characters are provided that may be used to visually represent the ASCII space character, U+2420 SYMBOL FOR SPACE, U+2422 BLANK SYMBOL, and U+2423 OPEN BOX.

***Standards.*** The CNS 11643 standard encodes characters for pictures of control codes. Standard representations for control characters have been defined—for example, in ANSI X3.32 and ISO 2047. If desired, the characters U+2400..U+241F may be used for these representations.

### ***Miscellaneous Technical: U+2300–U+23FF***

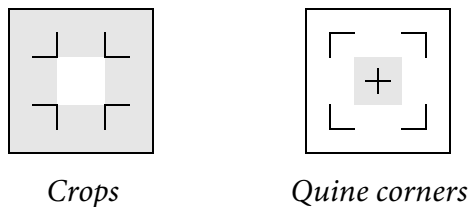
This block encodes technical symbols, including keytop labels such as U+232B ERASE TO THE LEFT. Excluded from consideration were symbols that are not normally used in one-dimensional text but are intended for two-dimensional diagrammatic use, such as most symbols for electronic circuits.

***Keytop Labels.*** Where possible, keytop labels have been unified with other symbols of like appearance—for example, U+21E7 UPWARDS WHITE ARROW to indicate the Shift key. While symbols such as U+2318 PLACE OF INTEREST SIGN and U+2388 HELM SYMBOL are generic symbols that have been adapted to use on keytops, other symbols specifically follow ISO/IEC 9995-7.

***Floor and Ceiling.*** The floor and ceiling symbols encoded at U+2308..U+230B are tall, narrow mathematical delimiters. These symbols should not be confused with the CJK corner brackets at U+300C and U+300D, which are wide characters used as quotation marks in East Asian text. They should also be distinguished from the half brackets at U+2E22..U+2E25, which are the most generally used editorial marks shaped like corner brackets. Additional types of editorial marks, including further corner bracket forms, can be found in the Supplemental Punctuation block (U+2E00..U+2E7F).

**Crops and Quine Corners.** Crops and quine corners are most properly used in two-dimensional layout but may be referred to in plain text. This usage is shown in Figure 22-9.

**Figure 22-9.** Usage of Crops and Quine Corners



**Angle Brackets.** U+2329 LEFT-POINTING ANGLE BRACKET and U+232A RIGHT-POINTING ANGLE BRACKET have long been canonically equivalent to the CJK punctuation characters U+3008 LEFT ANGLE BRACKET and U+3009 RIGHT ANGLE BRACKET, respectively. This canonical equivalence implies that the use of the latter (CJK) code points is preferred and that U+2329 and U+232A are also “wide” characters. (See Unicode Standard Annex #11, “East Asian Width,” for the definition of the East Asian wide property.) For this reason, the use of U+2329 and U+232A is deprecated for mathematics and for technical publication, where the wide property of the characters has the potential to interfere with the proper formatting of mathematical formulae. The angle brackets specifically provided for mathematics, U+27E8 MATHEMATICAL LEFT ANGLE BRACKET and U+27E9 MATHEMATICAL RIGHT ANGLE BRACKET, should be used instead. See Section 22.5, *Mathematical Symbols*.

**APL Functional Symbols.** APL (A Programming Language) makes extensive use of functional symbols constructed by composition with other, more primitive functional symbols. It used backspace and overstrike mechanisms in early computer implementations. In principle, functional composition is productive in APL; in practice, a relatively small number of composed functional symbols have become standard operators in APL. This relatively small set is encoded in its entirety in this block. All other APL extensions can be encoded by composition of other Unicode characters. For example, the APL symbol *a* underbar can be represented by U+0061 LATIN SMALL LETTER A + U+0332 COMBINING LOW LINE.

**Symbol Pieces.** The characters in the range U+239B..U+23B3, plus U+23B7, constitute a set of bracket and other symbol fragments for use in mathematical typesetting. These pieces originated in older font standards but have been used in past mathematical processing as characters in their own right to make up extra-tall glyphs for enclosing multiline mathematical formulae. Mathematical fences are ordinarily sized to the content that they enclose. However, in creating a large fence, the glyph is not scaled proportionally; in particular, the displayed stem weights must remain compatible with the accompanying smaller characters. Thus simple scaling of font outlines cannot be used to create tall brackets. Instead, a common technique is to build up the symbol from pieces. In particular, the characters U+239B LEFT PARENTHESIS UPPER HOOK through U+23B3 SUMMATION BOTTOM represent a set of glyph pieces for building up large versions of the fences (, ), [, ], {, and }, and of the large operators  $\Sigma$  and  $\int$ . These brace and operator pieces are compatibility char-

acters. They should not be used in stored mathematical text, although they are often used in the data stream created by display and print drivers.

Table 22-6 shows which pieces are intended to be used together to create specific symbols. For example, an instance of U+239B can be positioned relative to instances of U+239C and U+239D to form an extra-tall (three or more line) left parenthesis. The center sections encoded here are meant to be used only with the top and bottom pieces encoded adjacent to them because the segments are usually graphically constructed within the fonts so that they match perfectly when positioned at the same *x* coordinates.

**Table 22-6.** Use of Mathematical Symbol Pieces

|                   | Two-Row    | Three-Row        | Five-Row                     |
|-------------------|------------|------------------|------------------------------|
| Summation         | 23B2, 23B3 |                  |                              |
| Integral          | 2320, 2321 | 2320, 23AE, 2321 | 2320, 3×23AE, 2321           |
| Left parenthesis  | 239B, 239D | 239B, 239C, 239D | 239B, 3×239C, 239D           |
| Right parenthesis | 239E, 23A0 | 239E, 239F, 23A0 | 239E, 3×239F, 23A0           |
| Left bracket      | 23A1, 23A3 | 23A1, 23A2, 23A3 | 23A1, 3×23A2, 23A3           |
| Right bracket     | 23A4, 23A6 | 23A4, 23A5, 23A6 | 23A4, 3×23A5, 23A6           |
| Left brace        | 23B0, 23B1 | 23A7, 23A8, 23A9 | 23A7, 23AA, 23A8, 23AA, 23A9 |
| Right brace       | 23B1, 23B0 | 23AB, 23AC, 23AD | 23AB, 23AA, 23AC, 23AA, 23AD |

**Horizontal Brackets.** In mathematical equations, delimiters are often used horizontally, where they expand to the width of the expression they encompass. The six bracket characters in the range U+23DC..U+23E1 can be used for this purpose. In the context of mathematical layout, U+23B4 TOP SQUARE BRACKET and U+23B5 BOTTOM SQUARE BRACKET are also used that way. For more information, see Unicode Technical Report #25, “Unicode Support for Mathematics.”

The set of horizontal square brackets, U+23B4 TOP SQUARE BRACKET and U+23B5 BOTTOM SQUARE BRACKET, together with U+23B6 BOTTOM SQUARE BRACKET OVER TOP SQUARE BRACKET, are used by certain legacy applications to delimit vertical runs of text in non-CJK terminal emulation. U+23B6 is used where a single character cell is both the end of one such run and the start of another. The use of these characters in terminal emulation should not be confused with the use of rotated forms of brackets for vertically rendered CJK text. See the further discussion of this issue in *Section 6.2, General Punctuation*.

**Decimal Exponent Symbol.** U+23E8 DECIMAL EXPONENT SYMBOL is for compatibility with the Russian standard GOST 10859-64, as well as the paper tape and punch card standard, Alcor (DIN 66006). It represents a fixed token introducing the exponent of a real number in scientific notation, comparable to the more common usage of “e” in similar notations: 1.621e5. It was used in the early computer language ALGOL-60, and appeared in some Soviet-manufactured computers, such as the BESM-6 and its emulators. In the Unicode Standard it is treated simply as an atomic symbol; it is not considered to be equivalent to a generic subscripted form of the numeral “10” and is not given a decomposition. The vertical alignment of this symbol is slightly lower than the baseline, as shown in *Figure 22-10*.

**Figure 22-10.** Usage of the Decimal Exponent Symbol

```
СИСТЕМА АЛГОЛ-БЭСМ6. ВАРИАНТ 01-05-79.
СЧЕТ БЕЗ КОНТРОЛЯ
1. _BEGIN OUTPUT( 'Е' , 355.0/113.0) _END
-----
.314159292010+01
```

**Dental Symbols.** The set of symbols from U+23BE to U+23CC form a set of symbols from JIS X 0213 for use in dental notation.

**Metrical Symbols.** The symbols in the range U+23D1..U+23D9 are a set of spacing symbols used in the metrical analysis of poetry and lyrics.

**Electrotechnical Symbols.** The Miscellaneous Technical block also contains a smattering of electrotechnical symbols. These characters are not intended to constitute a complete encoding of all symbols used in electrical diagrams, but rather are compatibility characters encoded primarily for mapping to other standards. The symbols in the range U+238D..U+2394 are from the character set with the International Registry number 181. U+23DA EARTH GROUND and U+23DB FUSE are from HKSCS-2001.

**User Interface Symbols.** The characters U+231A, U+231B, and U+23E9 through U+23FA are often found in user interfaces for media players, clocks, alarms, and timers, as well as in text discussing those user interfaces. The black medium triangles (U+23F4..U+23F7) are the preferred shapes for User Interface purposes, rather than the similar geometric shapes located in the Geometric Shapes block: U+25A0..U+25FF. The Miscellaneous Symbols and Pictographs block also contains many user interface symbols in the ranges U+1F500..U+1F518, U+1F53A..U+1F53D and U+1F5BF..U+1F5DD, as well as clock face symbols in the range U+1F550..U+1F567.

**Standards.** This block contains a large number of symbols from ISO/IEC 9995-7:1994, *Information technology—Keyboard layouts for text and office systems—Part 7: Symbols used to represent functions*.

ISO/IEC 9995-7 contains many symbols that have been unified with existing and closely related symbols in Unicode. These symbols are shown with their ordinary shapes in the code charts, not with the particular glyph variation required by conformance to ISO/IEC 9995-7. Implementations wishing to be conformant to ISO/IEC 9995-7 in the depiction of these symbols should make use of a suitable font.

### **Optical Character Recognition: U+2440–U+245F**

This block includes those symbolic characters of the OCR-A character set that do not correspond to ASCII characters, as well as magnetic ink character recognition (MICR) symbols used in check processing.

**Standards.** Both sets of symbols are specified in ISO 2033.

### ***Symbols for Legacy Computing: U+1FB00-U+1FBFF***

The Unicode Standard encodes a number of symbols to support legacy computing graphic characters. Support for these legacy computing symbols includes 212 characters added in Version 13.0 to provide compatibility with a wide range of early home computers, or “microcomputers,” manufactured from the mid-1970s to the mid-1980s. These symbols also cover the teletext broadcasting standard originally developed in the early 1970s, and the Minitel standard developed in the 1980s. This collection of early microcomputer symbols includes support for the character sets of Amstrad CPC, Apple 8-bit, Atari 8 and 16-bit, Commodore 8 and 16-bit, MSX, Yamaha, RISC OS, and Tandy.

Most of the symbols in this block are semi-graphics: block-style symbols which can be combined to simulate an all-points-addressable graphic display. Many platforms used these semi-graphic characters to support a graphics mode: small blocks that would be plotted at various coordinates, resulting in the appropriate full-sized block character consisting of the necessary “on” and “off” blocks. Other symbols in the Symbols for Legacy Computing block include box drawing and shading characters, and miscellaneous arrows and stick figures. In the teletext specification, symbols in this group can be displayed either with cells joined together or with a narrow space between cells.

The Symbols for Legacy Computing block also includes clones of the ASCII digits 0 through 9 (U+1FBF0..U+1FBF9), styled as upright seven-segment digits that were often used in Atari 16-bit applications for game scores.

In addition to support for early microcomputers, teletext, and Minitel found in the Symbols for Legacy Computing block, terminal graphics legacy symbols are also encoded in the Miscellaneous Technical block. The legacy symbols in the Miscellaneous Technical block include block-style semi-graphics, border-colored characters, and box drawing characters. Other box drawing symbols are encoded in the Box Drawing block.

In particular, the Miscellaneous Technical block includes the horizontal scan line characters, U+23BA HORIZONTAL SCAN LINE-1 through U+23BD HORIZONTAL SCAN LINE-9, which represent characters that were encoded in character ROM for use with nine-line character graphic cells. Horizontal scan line characters are encoded for scan lines 1, 3, 7, and 9. The horizontal scan line character for scan line 5 is unified with U+2500 BOX DRAWINGS LIGHT HORIZONTAL.

The symbols in the Symbols for Legacy Computing block combined with the small number of vertical and horizontal line characters encoded in the Miscellaneous Technical block provide the compatibility characters needed for applications to emulate various early computer terminal support.

## 22.8 Geometrical Symbols

Geometrical symbols are a collection of geometric shapes and their derivatives plus block elements and characters used for box drawing in legacy environments. In addition to the blocks described in this section, the Miscellaneous Technical (U+2300..U+23FF), Miscellaneous Symbols (U+2600..U+26FF), and Miscellaneous Symbols and Arrows (U+2B00..U+2BFF) blocks contain geometrical symbols that complete the set of shapes in the Geometric Shapes block.

### *Box Drawing and Block Elements*

Box drawing and block element characters are graphic compatibility characters in the Unicode Standard. A number of existing national and vendor standards, including IBM PC Code Page 437, contain sets of characters intended to enable a simple kind of display cell graphics, assuming terminal-type screen displays of fixed-pitch character cells. The Unicode Standard does not encourage this kind of character-cell-based graphics model, but does include sets of such characters for backward compatibility with the existing standards.

**Box Drawing.** The Box Drawing block (U+2500..U+257F) contains a collection of graphic compatibility characters that originate in legacy standards in use prior to 1990 and that are intended for drawing boxes of various shapes and line widths for user interface components in character-cell-based graphic systems.

The “light,” “heavy,” and “double” attributes for some of these characters reflect the fact that the original sets often had a two-way distinction, between a light versus heavy line or a single versus double line, and included sufficient pieces to enable construction of graphic boxes with distinct styles that abutted each other in display.

In particular, the mappings to some Videotex mosaic drawing characters noted in the code charts refer to the concept of “heavy” as specified in early Videotex character registrations and Recommendations, which made a two-way distinction between light and heavy. See, for example, ITU-T Recommendation T.101, International Interworking for Videotex Services (November, 1988). The mappings do not reflect later Videotex registrations and modifications to the relevant Recommendations which specified three levels of weight distinction in lines for box drawing characters.

The lines in the box drawing characters typically extend to the middle of the top, bottom, left, and/or right of the bounding box for the character cell. They are designed to connect together into continuous lines, with no gaps between them. When emulating terminal applications, fonts that implement the box drawing characters should do likewise.

**Block Elements.** The Block Elements block (U+2580..U+259F) contains another collection of graphic compatibility characters. Unlike the box drawing characters, the legacy block elements are designed to fill some defined fraction of each display cell or to fill each display cell with some defined degree of shading. These elements were used to create crude graphic displays in terminals or in terminal modes on displays where bit-mapped graphics were unavailable.

Half-block fill characters are included for each half of a display cell, plus a graduated series of vertical and horizontal fractional fills based on one-eighth parts. The fractional fills do not form a logically complete set but are intended only for backward compatibility. There is also a set of quadrant fill characters, U+2596..U+259F, which are designed to complement the half-block fill characters and U+2588 FULL BLOCK. When emulating terminal applications, fonts that implement the block element characters should be designed so that adjacent glyphs for characters such as U+2588 FULL BLOCK create solid patterns with no gaps between them.

**Standards.** The box drawing and block element characters were derived from GB 2312, KS X 1001, a variety of industry standards, and several terminal graphics sets. The Videotex Mosaic characters, which have similar appearances and functions, are unified against these sets.

### ***Geometric Shapes: U+25A0–U+25FF***

The Geometric Shapes are a collection of characters intended to encode prototypes for various commonly used geometrical shapes—mostly squares, triangles, and circles. The collection is somewhat arbitrary in scope; it is a compendium of shapes from various character and glyph standards. The typical distinctions more systematically encoded include black versus white, large versus small, basic shape (square versus triangle versus circle), orientation, and top versus bottom or left versus right part.

**Hatched Squares.** The hatched and cross-hatched squares at U+25A4..U+25A9 are derived from the Korean national standard (KS X 1001), in which they were probably intended as representations of fill patterns. Because the semantics of those characters are insufficiently defined in that standard, the Unicode character encoding simply carries the glyphs themselves as geometric shapes to provide a mapping for the Korean standard.

**Lozenge.** U+25CA  $\diamond$  LOZENGE is a typographical symbol seen in PostScript and in the Macintosh character set. It should be distinguished from both the generic U+25C7 WHITE DIAMOND and the U+2662 WHITE DIAMOND SUIT, as well as from another character sometimes called a lozenge, U+2311 SQUARE LOZENGE.

**Use in Mathematics.** Many geometric shapes are used in mathematics. When used for this purpose, the center points of the glyphs representing geometrical shapes should line up at the center line of the mathematical font. This differs from the alignment used for some of the representative glyphs in the code charts.

For several simple geometrical shapes—circle, square, triangle, diamond, and lozenge—differences in size carry semantic distinctions in mathematical notation, such as the difference between use of the symbol as a variable or as one of a variety of operator types. The Miscellaneous Symbols and Arrows block contains numerous characters representing other sizes of these geometrical symbols. Several other blocks, such as General Punctuation, Mathematical Operators, Block Elements, Miscellaneous Symbols, and Geometric Shapes Extended, contain a few other characters which are members of the size-graded sets of such symbols.

For more details on the use of geometrical shapes in mathematics, see Unicode Technical Report #25, “Unicode Support for Mathematics.”

**Standards.** The Geometric Shapes are derived from a large range of national and vendor character standards. The squares and triangles at U+25E7..U+25EE are derived from the Linotype font collection. U+25EF LARGE CIRCLE is included for compatibility with the JIS X 0208-1990 Japanese standard.

**Geometric Shapes Extended: U+1F780–U+1F7FF**

The repertoire for the Geometric Shapes Extended block mostly originates from the set of Wingdings. It is intended primarily to complement existing sets of geometric shape symbols in other blocks. The choice of sizes for this extension is done with the goal that font designers will be able to scale uniformly among the various sizes for each class of geometric shapes. Table 22-7 provides a list of the sets that have characters spanning multiple blocks, including the Geometric Shapes Extended block. Differences in size may carry semantic distinction in mathematical notation.

**Table 22-7. Geometric Shape Collections**

| Description                                     | Code Points                                      |
|-------------------------------------------------|--------------------------------------------------|
| Black circles                                   | 22C5, 2219, 1F784, 2022, 2981, 26AB, 25CF, 2B24  |
| White circles                                   | 25CB, 2B58, 1F785..1F789                         |
| Colored circles                                 | 26AA, 26AB, 1F534, 1F535, 1F7E0..1F7E4           |
| Black squares                                   | 1F78C, 2B1D, 1F78D, 25AA, 25FE, 25FC, 25A0, 2B1B |
| White squares                                   | 25A1, 1F78E..1F792                               |
| Colored squares                                 | 2B1C, 2B1B, 1F7E5..1F7EB                         |
| White squares containing another black square   | 1F794, 25A3, 1F795                               |
| Black diamonds                                  | 1F797, 1F798, 2B29, 1F799, 2B25, 25C6            |
| White diamonds containing another black diamond | 1F79A, 25C8, 1F79B                               |
| Black lozenges                                  | 1F79D, 1F79E, 2B2A, 1F79F, 2B27, 29EB            |
| Five pointed stars                              | 1F7C9, 2605, 1F7CA, 272F                         |
| Six pointed stars                               | 2736, 1F7CB..1F7CD                               |
| Eight pointed stars                             | 2735, 1F7CE..1F7D1                               |
| Twelve pointed stars                            | 1F7D2, 2739, 1F7D3, 1F7D4                        |

This block also contains a set of colored circles and squares in the range U+1F7E0..U+1F7EB. Those colored circles and squares are intended for use with emoji, to augment the colored circles and other colored sets for emoji. Table 22-7 shows these sets, including white and black circles and squares, and red and blue circles from other blocks. Those sets are listed in the order: white, black, red, blue, orange, yellow, green, purple, brown. Unlike emoji modifiers for skin tone (see Unicode Technical Standard #51, “Unicode Emoji”), the symbols for colored circles and squares are simply graphical symbols which may convey the concepts of colors, but with no immediate implications for render-



ing of glyphs with those particular colors. For example, a user could specify a yellow circle symbol together with a ribbon emoji symbol to convey the notion of a “yellow ribbon,” but there would be no expectation that the font would combine the two characters and draw an actual yellow ribbon. These colored circles and squares are often used decoratively in emoji text, with no other semantic intent.

# 22.9 Miscellaneous Symbols

There are numerous blocks defined in the Unicode Standard which contain miscellaneous symbols that do not fit well into any of the categories of symbols already discussed. These include various small sets of special-use symbols such as zodiacal symbols, map symbols, symbols used in transportation and accommodation guides, dictionary symbols, gender symbols, and so forth. There are additional larger sets, such as sets of symbols for game pieces or playing cards, and divination symbols associated with the Yijing or other texts, as well as sets of medieval or ancient symbols used only in historical contexts.

Of particular note are the large number of pictographic symbols, called emoji (“picture character”), in common use on mobile devices. Many emoji originated from character sets, called carrier sets, in early widespread use in cell phones in Japan. A number of other symbols are commonly shown with an emoji-like presentation. The majority of such symbols are encoded in the blocks listed in *Table 22-8*, but many emoji characters are encoded in other blocks. For a complete listing of the historic carrier emoji sets, including information about which of those emoji characters have been unified with other symbol characters in the Unicode Standard, see the data file *EmojiSources.txt* in the Unicode Character Database. The list of all Unicode characters that normally may be candidates for emoji presentation can be found in the data file *emoji-data.txt* in the Unicode Character Database.

**Table 22-8.** Blocks with Characters Often Shown as Emoji

| Range        | Block Name                            |
|--------------|---------------------------------------|
| 2600..26FF   | Miscellaneous Symbols                 |
| 1F300..1F5FF | Miscellaneous Symbols and Pictographs |
| 1F600..1F64F | Emoticons                             |
| 1F680..1F6FF | Transport and Map Symbols             |
| 1F900..1F9FF | Supplemental Symbols and Pictographs  |
| 1FA70..1FAFF | Symbols and Pictographs Extended-A    |

An additional category of miscellaneous symbols are the so-called *dingbat* characters. These are essentially compatibility characters representing very specific glyph shapes associated with common “symbol” fonts in widespread legacy use. Symbols identified as “dingbats” are encoded in various blocks. The well-known “Zapf Dingbats” set is encoded comprehensively in the Dingbats block, U+2700..U+27BF. Other sets of dingbats, such as the Wingdings and Webdings sets, are encoded in various symbol blocks, but the majority are found in the Miscellaneous Symbols and Pictographs block, U+1F300..U+1F5FF.

Corporate logos and collections of graphical elements or pictures are not included in the Unicode Standard, because they tend either to be very specific in usage (logos, political party symbols, and so on) or are nonconventional in appearance and semantic interpretation (clip art collections), and hence are inappropriate for encoding as characters. The Unicode Standard recommends that such items be incorporated in text via higher-level protocols that allow intermixing of graphic images with text, rather than by indefinite extension of the number of miscellaneous symbols encoded as characters. Newer emoji-

like symbols using embedded graphics are already in widespread use on mobile phones and other devices.

**Rendering of Emoji.** Many of the characters in the blocks listed in *Table 22-8* are often presented in an emoji style. There may be a great deal of variability in presentation, along three axes:

- **Glyph shape:** Emoji may have a great deal of flexibility in the choice of glyph shape used to render them.
- **Color:** Many characters in an emoji context (such as cell phone e-mail or text messages) are displayed in color, sometimes as a multicolor image. While this is particularly true of emoji, there are other cases where non-emoji symbols, such as game symbols, may be displayed in color.
- **Animation:** Some characters in an emoji context are presented in animated form, usually as a repeating sequence of two to four images.

Emoji may be presented using color or animation, but need not be. Because many characters in the carrier emoji sets or other sources are unified with Unicode characters that originally came from other sources, it may not always be clear whether a character should be presented using an emoji style. However, for most such characters, variation sequences have been defined which can specify text or emoji presentation. Unicode Technical Standard #51, “Unicode Emoji,” provides some guidance about which characters should have which presentation style in various environments.

**Color Words in Unicode Character Names.** The representative glyph shown in the code charts for a character is always monochrome. The character name may include a term such as `BLACK` or `WHITE`, or in the case of characters for emoji pictographs, other color terms such as `BLUE` or `ORANGE`. The use of `BLACK` or `WHITE` in names such as `BLACK MEDIUM SQUARE` or `WHITE MEDIUM SQUARE` is generally intended to contrast filled versus outline shapes, or a darker color fill versus a lighter color fill; it is not intended to suggest that the character must be presented in black or white, respectively. Similarly, the color terms in names such as `BLUE HEART` or `ORANGE BOOK` are intended to help identify the characters; the characters may be presented using color, or in monochrome using different styles of shading or crosshatching, for example.

In Version 12.0 of the Unicode Standard, seven large, colored square emoji were added in the range `U+1F7E5..U+1F7EB`. Along with the earlier encoded `U+2B1B BLACK LARGE SQUARE` and `U+2B1C WHITE LARGE SQUARE`, these colored square emoji may be used in emoji ZWJ sequences to indicate that a base emoji should be displayed with the color of the square, if possible. The color of the square emoji is a general hint, and the color of the resulting image for the emoji ZWJ sequence need not be exactly the same as the colored square displayed by itself. Only a small number of such sequences are in the set of emoji sequences recommended for general interchange (RGI). See `emoji-zwj-sequences.txt`, documented in Annex A of Unicode Technical Standard #51, “Unicode Emoji.”

## Miscellaneous Symbols and Pictographs

The Miscellaneous Symbols (U+2600..U+26FF), Miscellaneous Symbols and Pictographs (U+1F300..U+1F5FF), Supplemental Symbols and Pictographs (U+1F900..U+1F9FF), and Symbols and Pictographs Extended-A (U+1FA70..U+1FAFF) blocks contain very heterogeneous collections of symbols that do not fit in any other Unicode character block and that tend to be pictographic in nature. These symbols are typically used for text decorations, but they may also be treated as normal text characters in applications such as typesetting chess books, card game manuals, and horoscopes.

The order of symbols in these blocks is arbitrary, but an attempt has been made to keep like symbols together and to group subsets of them into meaningful orders. Some of these subsets include weather and astronomical symbols, pointing hands, religious and ideological symbols, the Yijing (I Ching) trigrams, planet and zodiacal symbols, game symbols, musical dingbats, and recycling symbols. (For other moon phases, see the circle-based shapes in the Geometric Shapes block.)

**Standards.** The symbols in these blocks are derived from a large range of national and vendor character standards. Among them, characters from the Japanese Association of Radio Industries and Business (ARIB) standard STD-B24 are widely represented in the Miscellaneous Symbols block. The symbols from ARIB were initially used in the context of digital broadcasting, but in many cases their usage has evolved to more generic purposes. The Miscellaneous Symbols and Pictographs block includes many characters from the carrier emoji sets and the Wingdings/Webdings collections.

**Weather Symbols.** The characters in the ranges U+2600..U+2603, U+26C4..U+26CB, and U+1F321..U+1F32C, as well as U+2614 UMBRELLA WITH RAIN DROPS are weather symbols. These commonly occur as map symbols or in other contexts related to weather forecasting in digital broadcasting or on websites.

**Moon and Sun Symbols.** There are a variety of moon and sun symbols encoded in the Miscellaneous Symbols block (U+2609, U+263C..U+263E) and in the Miscellaneous Symbols and Pictographs block (U+1F311..U+1F31E). Some of these are used in astrological charts, while others are merely playful symbols showing faces. Various crescent signs for the moon do not necessarily represent particular phases of the moon.

The moon symbols in the range U+1F311..U+1F318, in particular, represent a systematic set of eight symbols for the phases of the moon. These symbols appear, for example, in moon charts, almanacs, tide tables, and similar documents to represent particular phases of the moon. There is a notable difference in interpretation of symbols for phases of the moon between Northern Hemisphere users and Southern Hemisphere users, with the graphical orientation of waxing and waning phases reversed. So, for example, in the Southern Hemisphere, U+1F312 WAXING CRESCENT MOON SYMBOL would usually be interpreted as representing the *waning* crescent moon, instead.

The use of these moon symbols (U+1F311..U+1F318) should follow the *shape* of the graphic symbols, as shown in the code charts. Users should not simply assume from the

character names that the symbols are intended to represent astronomical positions of the moon.

**Traffic Signs.** In general, traffic signs are quite diverse, tend to be elaborate in form and differ significantly between countries and locales. For the most part they are inappropriate for encoding as characters. However, there are a small number of conventional symbols which have been used as characters in contexts such as digital broadcasting or mobile phones. The characters in the ranges U+26CC..U+26CD and U+26CF..U+26E1 are traffic sign symbols of this sort, encoded for use in digital broadcasting. Additional traffic signs are included in the Transport and Map Symbols block.

**Dictionary and Map Symbols.** The characters in the range U+26E8..U+26FF are dictionary and map symbols used in the context of digital broadcasting. Numerous other symbols in this block and scattered in other blocks also have conventional uses as dictionary or map symbols. For example, these may indicate special uses for words, or indicate types of buildings, points of interest, particular activities or sports, and so on.

**Plastic Bottle Material Code System.** The seven numbered logos encoded from U+2673 to U+2679, ♻️♻️♻️♻️♻️♻️♻️, are from “The Plastic Bottle Material Code System,” which was introduced in 1988 by the Society of the Plastics Industry (SPI). This set consistently uses thin, two-dimensional curved arrows suitable for use in plastics molding. In actual use, the symbols often are combined with an abbreviation of the material class below the triangle. Such abbreviations are not universal; therefore, they are not present in the representative glyphs in the code charts.

**Recycling Symbol for Generic Materials.** An unnumbered plastic resin code symbol U+267A ♻️ RECYCLING SYMBOL FOR GENERIC MATERIALS is not formally part of the SPI system but is found in many fonts. Occasional use of this symbol as a generic materials code symbol can be found in the field, usually with a text legend below, but sometimes also surrounding or overlaid by other text or symbols. Sometimes the UNIVERSAL RECYCLING SYMBOL is substituted for the generic symbol in this context.

**Universal Recycling Symbol.** The Unicode Standard encodes two common glyph variants of this symbol: U+2672 ♻️ UNIVERSAL RECYCLING SYMBOL and U+267B ♻️ BLACK UNIVERSAL RECYCLING SYMBOL. Both are used to indicate that the material is recyclable. The white form is the traditional version of the symbol, but the black form is sometimes substituted, presumably because the thin outlines of the white form do not always reproduce well.

**Paper Recycling Symbols.** The two paper recycling symbols, U+267C ♻️ RECYCLED PAPER SYMBOL and U+267D ♻️ PARTIALLY-RECYCLED PAPER SYMBOL, can be used to distinguish between fully and partially recycled fiber content in paper products or packaging. They are usually accompanied by additional text.

**Gender Symbols.** The characters in the range U+26A2..U+26A9 are gender symbols. These are part of a set with U+2640 FEMALE SIGN, U+2642 MALE SIGN, U+26AA MEDIUM WHITE CIRCLE, and U+26B2 NEUTER. They are used in sexual studies and biology, for example. Some of these symbols have other uses as well, as astrological or alchemical symbols.

**Genealogical Symbols.** The characters in the range U+26AD..U+26B1 are sometimes seen in genealogical tables, where they indicate marriage and burial status. They may be augmented by other symbols, including the small circle indicating betrothal.

**Game Symbols.** The Miscellaneous Symbols block also contains a variety of small symbol sets intended for the representation of common game symbols or tokens in text. These include symbols for playing card suits, often seen in manuals for bridge and other card games, as well as a set of dice symbols. The chess symbols are often seen in figurine algebraic notation. In addition, there are symbols for game pieces or notation markers for go, shogi (Japanese chess), and draughts (checkers).

Larger sets of game symbols are encoded in their own blocks. See the discussion of playing cards, chess symbols, mahjong tile symbols, and domino tile symbols later in this section.

**Animal Symbols.** The animal symbol characters in the range U+1F400..U+1F42C are encoded primarily to cover the emoji sets used by Japanese cell phone carriers. Animal symbols are widely used in Asia as signs of the zodiac, and that is part of the reason for their inclusion in the cell phone sets. However, the particular animal symbols seen in Japan and China are not the only animals used as zodiacal symbols throughout Asia. The set of animal symbols encoded in this block includes other animal symbols used as zodiacal symbols in Vietnam, Thailand, Persia, and other Asian countries. These zodiacal uses are specifically annotated in the Unicode code charts.

Other animal symbols have no zodiacal associations, and are included simply to cover the carrier emoji sets. A few of the animal symbols have conventional uses to designate types of meat on menus. Later additions of animal symbols fill perceived gaps in the set, responding to the wide popularity of animal symbols in Unicode-based emoji implementations.

**Cultural Symbols.** The five cultural symbols encoded in the range U+1F5FB..U+1F5FF mostly designate cultural landmarks of particular importance to Japan. They are encoded for compatibility with emoji sets used by Japanese cell phone carriers, and are not intended to set a precedent for encoding additional sets of cultural landmarks or other pictographic cultural symbols as characters.

**Hand Symbols.** The pictographic symbols for hands encoded in the ranges U+1F90F, U+1F918..U+1F91F, U+1F446..U+1F450, and U+1F58E..U+1F5A3, as well as in the U+270A..U+270D range in the Dingbats block, represent various hand gestures. The interpretations associated with such gestures vary significantly among cultures.

**Emoji Modifiers.** The emoji modifiers U+1F3FB..U+1F3FF designate five different skin tones based on the Fitzpatrick scale. These may be displayed in isolation as color or half-tone swatches, or they may form a ligature with a preceding emoji character representing a person or body part in order to specify a particular appearance for that character.

**Miscellaneous Symbols in Other Blocks.** In addition to the blocks described in this section, which are devoted entirely to sets of miscellaneous symbols, there are many other blocks which contain small numbers of otherwise uncategorized symbols. See, for example, the Miscellaneous Symbols and Arrows block U+2B00..U+2B7F, the Enclosed Alphanumeric Supplement block U+1F100..U+1F1FF, the CJK Symbols and Punctuation block

U+3000..U+303F, and the Ideographic Symbols and Punctuation block 16FE0..16FFF. Some of these blocks contain symbols which extend or complement sets of symbols contained in the Miscellaneous Symbols block.

### ***Emoticons: U+1F600–U+1F64F***

Emoticons (from “emotion” plus “icon”) originated as a way to convey emotion or attitude in e-mail messages, using ASCII character combinations such as :- ) to indicate a smile—and by extension, a joke—and :-( to indicate a frown. In East Asia, a number of more elaborate sequences have been developed, such as (")(-\_-)(") showing an upset face with hands raised.

Over time, many systems began replacing such sequences with images, and also began providing a way to input emoticon images directly, such as a menu or palette. The carrier emoji sets used by Japanese cell phone providers contain a large number of characters for emoticon images, and most of the characters in this block are from those sets. They are divided into a set of humanlike faces, a smaller set of cat faces that parallel some of the humanlike faces, and a set of gesture symbols that combine a human or monkey face with arm and hand positions.

Several emoticons are also encoded in the Miscellaneous Symbols block at U+2639..U+263B and in the Supplemental Symbols and Pictographs block at U+1F910..U+1F917 and U+1F920..1F927.

### ***Transport and Map Symbols: U+1F680–U+1F6FF***

This block is similar to the blocks Miscellaneous Symbols and Miscellaneous Symbols and Pictographs, but is a more cohesive set of symbols. Many of these symbols originated in the emoji sets used by Japanese cell phone carriers.

Various traffic signs and map symbols are also encoded in the Miscellaneous Symbols block.

### ***Dingbats: U+2700–U+27BF***

Most of the characters in the Dingbats block are derived from a well-established set of glyphs, the ITC Zapf Dingbats series 100, which constitutes the industry standard “Zapf Dingbat” font currently available in most laser printers. The order of the Dingbats block basically follows the PostScript encoding. Dingbat characters derived from the Wingdings and Webdings sets are encoded in other blocks, particularly in the Miscellaneous Symbols and Pictographs block, U+1F300..U+1F5FF.

**Unifications and Additions.** Where a dingbat from the ITC Zapf Dingbats series 100 could be unified with a generic symbol widely used in other contexts, only the generic symbol was encoded. Examples of such unifications include card suits, BLACK STAR, BLACK TELEPHONE, and BLACK RIGHT-POINTING INDEX (see the Miscellaneous Symbols block); BLACK CIRCLE and BLACK SQUARE (see the Geometric Shapes block); white encircled numbers 1 to

10 (see the Enclosed Alphanumerics block); and several generic arrows (see the Arrows block). Those four entries appear elsewhere in this chapter. Other dingbat-like characters, primarily from the carrier emoji sets, are encoded in the gaps that resulted from this unification.

In other instances, other glyphs from the ITC Zapf Dingbats series 100 glyphs have come to be recognized as having applicability as generic symbols, despite having originally been encoded in the Dingbats block. For example, the series of negative (black) circled numbers 1 to 10 are now treated as generic symbols for this sequence, the continuation of which can be found in the Enclosed Alphanumerics block. Other examples include U+2708 AIRPLANE and U+2709 ENVELOPE, which have definite semantics independent of the specific glyph shape, and which therefore should be considered generic symbols rather than symbols representing only the Zapf Dingbats glyph shapes.

For many of the remaining characters in the Dingbats block, their semantic value is primarily their shape; unlike characters that represent letters from a script, there is no well-established range of typeface variations for a dingbat that will retain its identity and therefore its semantics. It would be incorrect to arbitrarily replace U+279D TRIANGLE-HEADED RIGHTWARDS ARROW with any other right arrow dingbat or with any of the generic arrows from the Arrows block (U+2190..U+21FF). However, exact shape retention for the glyphs is not always required to maintain the relevant distinctions. For example, ornamental characters such as U+2741 EIGHT PETALLED OUTLINED BLACK FLORETTE have been successfully implemented in font faces other than Zapf Dingbats with glyph shapes that are similar, but not identical to the ITC Zapf Dingbats series 100.

The following guidelines are provided for font developers wishing to support this block of characters. Characters showing large sets of contrastive glyph shapes in the Dingbats block, and in particular the various arrow shapes at U+2794..U+27BE, should have glyphs that are closely modeled on the ITC Zapf Dingbats series 100, which are shown as representative glyphs in the code charts. The same applies to the various stars, asterisks, snowflakes, drop-shadowed squares, check marks, and x's, many of which are ornamental and have elaborate names describing their glyphs.

Where the preceding guidelines do not apply, or where dingbats have more generic applicability as symbols, their glyphs do not need to match the representative glyphs in the code charts in every detail.

**Ornamental Brackets.** The 14 ornamental brackets encoded at U+2768..U+2775 are part of the set of Zapf Dingbats. Although they have always been included in Zapf Dingbats fonts, they were unencoded in PostScript versions of the fonts on some platforms. The Unicode Standard treats these brackets as punctuation characters.

### ***Ornamental Dingbats: U+1F650–U+1F67F***

The block contains a variety of text ornaments and ornamental punctuation marks similar to characters encoded in the main Dingbats block. Most of these symbols are encoded for compatibility with Wingdings or Wingdings 2 font usage; a few derive from Webdings.



Many of these dingbats can be classified as fleurons. These constitute variations on the theme represented by the classic heder a or ivy leaf shape encoded as U+2767 ROTATED FLO-RAL HEART BULLET.

The block also contains stylistic variants of punctuation marks, including numerous styles of ampersands and et-ligatures, quotation marks, and question marks. These characters extend similar sets of stylized punctuation marks in the Dingbats block. All of these stylized ornamental variants are treated as symbols rather than as true punctuation in the standard.

### ***Alchemical Symbols: U+1F700–U+1F77F***

Alchemical symbols were first used by Greek, Syriac, and Egyptian writers around the fifth or sixth century CE and were adopted and proliferated by medieval Arabic and European writers. European alchemists, natural philosophers, chemists, and apothecaries developed and used several parallel systems of symbols while retaining many symbols created by Greek, Syriac, and medieval Arabic writers. Alchemical works published in what is best described as a textbook tradition in the seventeenth and eighteenth centuries routinely included tables of symbols that probably served to spread their use. They became obsolete as alchemy gave way to chemistry. Nevertheless, alchemical symbols continue to be used extensively today in scholarly literature, creative works, New Age texts, and in the gaming and graphics industries.

This block contains a core repertoire of symbols recognized and organized into tables by European writers working in the alchemical textbook tradition approximately 1620–1720. This core repertoire includes all symbols found in the vast majority of the alchemical works of major figures such as Newton, Boyle, and Paracelsus. Some of the most common alchemical symbols have multiple meanings, and are encoded in the Miscellaneous Symbols block, where their usage as alchemical symbols is annotated. For example, U+2609 SUN is also an alchemical symbol for gold.

The character names for the alchemical symbols are in English. Their equivalent Latin names, which often were in greater currency during the period of greatest use of these symbols, are provided as aliases in the code charts. Some alchemical names in English directly derive from the Latin name, such as aquafortis and aqua regia, so in a number of cases the English and Latin names are identical.

### ***Mahjong Tiles: U+1F000–U+1F02F***

The characters in this block are game symbols representing the set of tiles used to play the popular Chinese game of mahjong. The exact origin of mahjong is unknown, but it has been around since at least the mid-19th century, and its popularity spread to Japan, Britain, and the United States during the early 20th century.

Like other game symbols in the Unicode Standard, the mahjong tile symbols are intended as abstractions of graphical symbols for game pieces used in text. Simplified, iconic representation of mahjong pieces are printed in game manuals and appear in discussion about the game. There is some variation in the exact set of tiles used in different countries, so the

Unicode Standard encodes a superset of the graphical symbols for the tiles used in the various local traditions. The main set of tiles consists of three suits with nine numerical tiles each: the Bamboos, the Circles, and the Characters.

Additional tiles include the Dragons, the Winds, the Flowers, and the Seasons. The blank tile symbol is the so-called *white dragon*. Also included is a black tile symbol, which does not represent an actual game tile, but rather indicates a facedown tile, occasionally seen as a symbol in text about playing mahjong.

### ***Domino Tiles: U+1F030–U+1F09F***

This block contains a set of graphical symbols for domino tiles. Dominoes is a game which derives from Chinese tile games dating back to the twelfth century.

Domino tile symbols are used for the “double-six” set of tiles, which is the most common set of dominoes and the only one widely attested in manuals and textual discussion using graphical tile symbols.

The domino tile symbols do not represent the domino pieces per se, but instead constitute graphical symbols for particular orientations of the dominoes, because orientation of the tiles is significant in discussion of dominoes play. Each visually distinct rotation of a domino tile is separately encoded. Thus, for example, both U+1F081 DOMINO TILE VERTICAL-04-02 and U+1F04F DOMINO TILE HORIZONTAL-04-02 are encoded, as well as U+1F075 DOMINO TILE VERTICAL-02-04 and U+1F043 DOMINO TILE HORIZONTAL-02-04. All four of those symbols represent the same game tile, but each orientation of the tile is visually distinct and requires its own symbol for text. The digits in the character names for the domino tile symbols reflect the dot patterns on the tiles.

Two symbols do not represent particular tiles of the double-six set of dominoes, but instead are graphical symbols for a domino tile turned facedown.

### ***Playing Cards: U+1F0A0–U+1F0FF***

The symbols in this block are used to represent the 52-card deck most commonly used today, and the 56-card deck used in some European games; the latter includes a Knight in addition to Jack, Queen, and King. These cards map completely to the Minor Arcana of the Western Tarot from which they derive, and are unified with the latter. The symbols for trumps in the range U+1F0E0..U+1F0F5 occur as playing cards in some traditional German, Italian, and French decks. These trumps are historically derived from the 22 Major Arcana of the esoteric Western Tarot sets. The combined set can be used to represent the 78 cards of the common tarot decks.

Also included in this block are a generic card back and three jokers. U+1F0CF PLAYING CARD BLACK JOKER is used in one of the Japanese cell phone carrier emoji sets; its presentation may be in color and need not be black. U+1F0BF PLAYING CARD RED JOKER occurs in some card decks as a third joker.

These characters most commonly appear as the Anglo-French-style playing cards used with international bridge or poker. However, playing card characters may have a variety of different appearances depending on language and usage. In different countries, the suits, colors and numbers may be substantially different, to the point of being unrecognizable. For example, the letters on face cards may vary (English cards use “K” for “king,” while French cards use “R” for “roi”); the digits on the numbered cards may appear as a Western “10” or as “१०” in Hindi, and the appearance of the suits may differ (Swiss playing cards depict acorns rather than clubs, while traditional tarot cards use swords rather than spades). The background decoration of cards may also vary radically. When used to represent the cards of divination tarot decks, the visual appearance is usually very different and much more complex.

No one should expect reliable interchange of a particular appearance of the playing card characters without additional information (such as a font) or agreement between sender and receiver. Without such information or agreement, someone viewing an online document may see substantially different glyphs from what the writer intended.

Basic playing card suit symbols are encoded in the Miscellaneous Symbols block in the range U+2660..U+2667.

### ***Chess Symbols: U+1FA00–U+1FA6F***

The Chess Symbols block contains extensions for chess notations beyond the basic Western chess symbols found in the Miscellaneous Symbols block. The chess symbols in the range U+1FA00..U+1FA53 are used in a variety of heterodox Western chess notations, also widely referred to as “fairy chess.” These notations include the introduction of new or hybrid chess pieces, such as grasshoppers, nightriders, equihoppers, or various blends of knights with other pieces. There are also a number of neutral pieces, which conceptually belong neither to the white side nor the black side, often displayed with one side of the piece black and the other side of the piece shown with an outlined glyph. Many of these symbols simply consist of existing Western chess symbols for orthodox pieces, inverted or turned sideways. This practice dates from the time when printers would often take existing cast metal sorts and physically invert or turn them before locking them into the forme, to create new symbols for printing heterodox chess problems and commentary.

This block also contains a set of circled CJK ideographic symbols used in Chinese chess (*Xiangqi*) notation, in the range U+1FA60..U+1FA6D. These symbols come in separate “red” and “black” sets, abstractly representing the two sets of seven pieces in that game. In actual practice, both for the symbols printed on the pieces in Chinese chess sets and in notation, there is considerable variation in the color of the pieces, and in the particular CJK ideograph within the circle. For example, both traditional and simplified characters occur, and there is some other variation in the choice of the CJK ideograph, as well. Because of this variability in the CJK ideograph used, these symbols are treated differently than most regular circled CJK ideographic symbols in the standard. No compatibility decompositions to CJK unified ideographs are given in the UCD or shown in the code charts.

## ***Yijing Hexagram Symbols: U+4DC0–U+4DF6***

Usage of the Yijing Hexagram Symbols in China begins with a text called 《周易》 *Zhou Yi*, (“the Zhou Dynasty classic of change”), said to have originated circa 1000 BCE. This text is now popularly known as the *Yijing*, *I Ching*, or *Book of Changes*. These symbols represent a primary level of notation in this ancient philosophical text, which is traditionally considered the first and most important of the Chinese classics. Today, these symbols appear in many print and electronic publications, produced in Asia and all over the world. The important Chinese character lexicon *Hanyu Da Zidian*, for example, makes use of these symbols in running text. These symbols are semantically distinct written signs associated with specific words. Each of the 64 hexagrams has a unique one- or two-syllable name. Each hexagram name is intimately connected with interpretation of the six lines. Related characters are Monogram and Digram Symbols (U+268A..U+268F), Yijing Trigram Symbols (U+2630..U+2637), and Tai Xuan Jing Symbols (U+1D300..U+1D356).

## ***Tai Xuan Jing Symbols: U+1D300–U+1D356***

Usage of these symbols in China begins with a text called 《太玄經》 *Tai Xuan Jing* (literally, “the exceedingly arcane classic”). Composed by a man named 楊雄 *Yang Xiong* (53 BCE–18 CE), the first draft of this work was completed in 2 BCE, in the decade before the fall of the Western Han Dynasty. This text is popularly known in the West under several titles, including *The Alternative I Ching* and *The Elemental Changes*. A number of annotated editions of *Tai Xuan Jing* have been published and reprinted in the 2,000 years since the original work appeared.

These symbols represent a primary level of notation in the original ancient text, following and expanding upon the traditions of the Chinese classic *Yijing*. The tetragram signs are less well known and less widely used than the hexagram signs. For this reason they were encoded on Plane 1 rather than the BMP.

**Monograms.** U+1D300 MONOGRAM FOR EARTH is an extension of the traditional Yijing monogram symbols, U+268A MONOGRAM FOR YANG and U+268B MONOGRAM FOR YIN. Because *yang* is typically associated with heaven (Chinese *tian*) and *yin* is typically associated with earth (Chinese *di*), the character U+1D300 has an unfortunate name. Tai Xuan Jing studies typically associate it with human (Chinese *ren*), as midway between heaven and earth.

**Digrams.** The range of characters U+1D301..U+1D302 constitutes an extension of the Yijing digram symbols encoded in the range U+268C..U+268F. They consist of the combinations of the human (*ren*) monogram with either the *yang* or the *yin* monogram. Because of the naming problem for U+1D300, these digrams also have infelicitous character names. Users are advised to identify the digram symbols by their representative glyphs or by the Chinese aliases provided for them in the code charts.

**Tetragrams.** The bulk of the symbols in the Tai Xuan Jing Symbols block are the tetragram signs. These tetragram symbols are semantically distinct written signs associated with specific words. Each of the 81 tetragrams has a unique monosyllabic name, and each tetragram name is intimately connected with interpretation of the four lines.

The 81 tetragram symbols (U+1D306..U+1D356) encoded on Plane 1 constitute a complete set. Within this set of 81 signs, a subset of 16 signs known as the Yijing tetragrams is of importance to Yijing scholarship. These are used in the study of the “nuclear trigrams.” Related characters are the Yijing Trigram symbols (U+2630..U+2637) and the Yijing Hexagram symbols (U+4DC0..U+4DFF).

### ***Ancient Symbols: U+10190–U+101CF***

This block contains ancient symbols, none of which are in modern use. Typically, they derive from ancient epigraphic, papyrological, or manuscript traditions, and represent miscellaneous symbols not specifically included in blocks dedicated to particular ancient scripts. The first set of these consists of ancient Roman symbols for weights and measures, and symbols used in Roman coinage.

Similar symbols can be found in the Ancient Greek Numbers block, U+10140..U+1018F.

### ***Phaistos Disc Symbols: U+101D0–U+101FF***

The Phaistos disc was found during an archaeological dig in Phaistos, Crete about a century ago. The small fired clay disc is imprinted on both sides with a series of symbols, arranged in a spiral pattern. The disc probably dates from the mid-18th to the mid-14th century BCE.

The symbols have not been deciphered, and the disc remains the only known example of these symbols. Because there is nothing to compare them to, and the corpus is so limited, it is not even clear whether the symbols constitute a writing system for a language or are something else entirely. Nonetheless, the disc has engendered great interest, and numerous scholars and amateurs spend time discussing the symbols.

The repertoire of symbols is noncontroversial, as they were incised in the disc by stamping preformed seals into the clay. Most of the symbols are clearly pictographic in form. The entire set is encoded in the Phaistos Disc Symbols block as a set of symbols, with no assumptions about their possible meaning and functions. One combining mark is encoded. It represents a hand-carved mark on the disc, which occurs attached to the final sign of groups of other symbols.

***Directionality.*** Scholarly consensus is that the text of the Phaistos disc was inscribed starting from the outer rim of the disc and going inward toward the center. Because of that layout order and the orientation of the spiral, the disc text can be said to have right-to-left directionality. However, the Phaistos disc symbols have been given a default directionality of strong left-to-right in the Unicode Standard. This choice simplifies text layout of the symbols for researchers and would-be decipherers, who wish to display the symbols in the same order as the surrounding left-to-right text (for example, in the Latin script) used to discuss them. The additional complexity of bidirectional layout and editing would be unwelcome in such contexts.

This choice of directionality properties for the Phaistos disc symbols matches the precedent of the Old Italic script. (See *Section 8.6, Old Italic.*) Early Old Italic inscriptions were

often laid out from right to left, but the directionality of the Old Italic script in the Unicode Standard is strong-left-to-right, to simplify layout using the modern scholarly conventions for discussion of Old Italic texts.

The glyphs for letters of ancient Mediterranean scripts often show mirroring based on line direction. This behavior is well-known, for example, for archaic Greek when written in boustrophedon. Etruscan also displays glyph mirroring of letters. The choice of representative glyphs for the Phaistos disc symbols is based on this mirroring convention, as well. The symbols on the disc are in a right-to-left line context. However, the symbols are given left-to-right directionality in the Unicode Standard, so the representative glyphs in the code charts are reversed (mirrored) from their appearance on the disc.

## 22.10 Enclosed and Square

There are a large number of compatibility symbols in the Unicode Standard which consist either of letters or numbers enclosed in some graphic element, or which consist of letters or numbers in a square arrangement. Many of these symbols are derived from legacy East Asian character sets, in which such symbols are commonly encoded as elements.

**Enclosed Symbols.** Enclosed symbols typically consist of a letter, digit, Katakana syllable, Hangul jamo, or CJK ideograph enclosed in a circle or a square. In some cases the enclosure may consist of a pair of parentheses or tortoise-shell brackets, and the enclosed element may also consist of more than a single letter or digit, as for circled numbers 10 through 50. Occasionally the symbol is shown as white on a black encircling background, in which case the character name typically includes the word `NEGATIVE`.

Many of the enclosed symbols that come in small, ordered sets—the Latin alphabet, kana, jamo, digits, and Han ideographs one through ten—were originally intended for use in text as numbered bullets for lists. Parenthetical enclosures were in turn developed to mimic typewriter conventions for representing circled letters and digits used as list bullets. This functionality has now largely been supplanted by styles and other markup in rich text contexts, but the enclosed symbols in the Unicode Standard are encoded for interoperability with the legacy East Asian character sets and for the occasional text context where such symbols otherwise occur.

A few of the enclosed symbols have conventional meanings unrelated to the usage of encircled letters and digits as list bullets. In some instances these are distinguished in the standard—often because legacy standards separately encoded them. Thus, for example, U+24B8 © CIRCLED LATIN CAPITAL LETTER C is distinct from U+00A9 © COPYRIGHT SIGN, even though the two symbols are similar in appearance. In cases where otherwise generic enclosed symbols have specific conventional meanings, those meanings are called out in the code charts with aliases or other annotations. For example, U+1F157 ⑤ NEGATIVE CIRCLED LATIN CAPITAL LETTER H is also a commonly occurring map symbol for “hotel.”

**Square Symbols.** Another convention commonly seen in East Asian character sets is the creation of compound symbols by arranging two, three, four, or even more small-sized letters or syllables into a square shape consistent with the typical rendering footprint of a CJK ideograph. One subset of these consists of square symbols of square symbols for Latin abbreviations, often for SI and other technical units, such as “km” or “km/h”; these square symbols are mostly derived from Korean legacy standards. Another subset consists of Katakana words for units of measurement, classified ad symbols, and many other similar word elements arranged into a square array; these symbols are derived from Japanese legacy standards. A third major subset consists of Chinese telegraphic symbols for hours, days, and months, consisting of a digit or sequence of digits next to the CJK ideograph for “hour,” “day” or “month.”

**Source Standards.** Major sources for the repertoire of enclosed and square symbols in the Unicode Standard include the Korean national standard, KS X 1001:1998; the Chinese national standard, GB 2312:1980; the Japanese national standards JIS X 0208-1997 and JIS X 0213:2000; and CNS 11643. Others derive from the Japanese television standard, ARIB

STD B24, and from various East Asian industry standards, such as the Japanese cell phone carrier emoji sets, or corporate glyph registries.

**Allocation.** The Unicode Standard includes five blocks allocated for the encoding of various enclosed and square symbols. Each of those blocks is described briefly in the text that follows, to indicate which subsets of these symbols it contains and to highlight any other special considerations that may apply to each block. In addition, there are a number of circled digit and number symbols encoded in the Dingbats block (U+2700..U+27BF). Those circled symbols occur in the ITC Zapf dingbats series 100, and most of them were encoded with other Zapf dingbat symbols, rather than being allocated in the separate blocks for enclosed and square symbols. Finally, a small number of circled symbols from ISO/IEC 8859-1 or other sources can be found in the Latin-1 Supplement block (U+0080..U+00FF) or the Letterlike Symbols block (U+2100..U+214F).

**Decomposition.** Nearly all of the enclosed and square symbols in the Unicode Standard are considered compatibility characters, encoded for interoperability with other character sets. A significant majority of those are also compatibility decomposable characters, given explicit compatibility decompositions in the Unicode Character Database. The general patterns for these decompositions are described here. For full details for any particular one of these symbols, see the code charts or consult the data files in the UCD.

Parenthesized symbols are decomposed to sequences of opening and closing parentheses surrounding the letter or digit(s) of the symbol. Square symbols consisting of digit(s) followed by a full stop or a comma are decomposed into the digit sequence and the full stop or comma. Square symbols consisting of several Katakana syllables are decomposed into the corresponding sequence of Katakana characters and are given the decomposition tag “<square>”. Similar principles apply to square symbols consisting of sequences of Latin letters and symbols. Chinese telegraphic symbols, consisting of sequences of digits and CJK ideographs, are given compatibility decompositions, but do not have the decomposition tag “<square>”.

Circled symbols consisting of a single letter or digit surrounded by a simple circular graphic element are given compatibility decompositions with the decomposition tag “<circle>”. Circled symbols with more complex graphic styles, including double circled and negative circled symbols, are simply treated as atomic symbols, and are not decomposed. The same pattern is applied to enclosed symbols where the enclosure is a square graphic element instead of a circle, except that the decomposition tag in those cases is “<square>”. Occasionally a “circled” symbol that involves a sequence of Latin letters is preferentially represented with an ellipse surrounding the letters, as for U+1F12E  CIRCLED WZ, the German *Warenzeichen*. Such elliptic shape is considered to be a typographical adaptation of the circle, and does not constitute a distinct decomposition type in the Unicode Standard.

It is important to realize that the decomposition of enclosed symbols in the Unicode Standard does not make them canonical equivalents to letters or digits in sequence with combining enclosing marks such as U+20DD  COMBINING ENCLOSING CIRCLE. The combining enclosing marks are provided in the Unicode Standard to enable the representation of occasional enclosed symbols not otherwise encoded as characters. There is also



no defined way of indicating the application of a combining enclosing mark to more than a single base character. Furthermore, full rendering support of the application of enclosing combining marks, even to single base characters, is not widely available. Hence, in most instances, if an enclosed symbol is available in the Unicode Standard as a single encoded character, it is recommended to simply make use of that composed symbol.

**Casing.** There are special considerations for the casing relationships of enclosed or square symbols involving letters of the Latin alphabet. The *circled* letters of the Latin alphabet come in an uppercase set (U+24B6..U+24CF) and a lowercase set (U+24D0..U+24EA). Largely because the compatibility decompositions for those symbols are to a single letter each, these two sets are given the derived properties, Uppercase and Lowercase, respectively, and case map to each other. The superficially similar *parenthesized* letters of the Latin alphabet also come in an uppercase set (U+1F110..U+1F129) and a lowercase set (U+24BC..U+24B5), but are not case mapped to each other and are not given derived casing properties. This difference is in part because the compatibility decompositions for these parenthesized symbols are to sequences involving parentheses, instead of single letters, and in part because the uppercase set was encoded many years later than the lowercase set. Square symbols consisting of arbitrary sequences of Latin letters, which themselves may be of mixed case, are simply treated as caseless symbols in the Unicode Standard.

### ***Enclosed Alphanumerics: U+2460–U+24FF***

The enclosed symbols in this block consist of single Latin letters, digits, or numbers—most enclosed by a circle. The block also contains letters, digits, or numbers enclosed in parentheses, and a series of numbers followed by full stop. All of these symbols are intended to function as numbered (or lettered) bullets in ordered lists, and most are encoded for compatibility with major East Asian character sets.

The circled numbers one through ten (U+2461..U+2469) are also considered to be unified with the comparable set of circled black numbers with serifs on a white background from the ITC Zapf Dingbats series 100. Those ten symbols are encoded in this block, instead of in the Dingbats block.

The negative circled numbers eleven through twenty (U+24EB..U+24F4) are a continuation of the set of circled white numbers with serifs on a black background, encoded at U+2776..U+277F in the Dingbats block.

### ***Enclosed CJK Letters and Months: U+3200–U+32FF***

This block contains large sets of circled or parenthesized Japanese Katakana, Hangul jamo, or CJK ideographs, from East Asian character sets. It also contains circled numbers twenty-one through fifty, which constitute a continuation of the series of circled numbers from the Enclosed Alphanumerics block. There are also a small number of Chinese telegraph symbols and square Latin abbreviations, which are continuations of the larger sets primarily encoded in the CJK Compatibility block.

The enclosed symbols in the range U+3248..U+324F, which consist of circled numbers ten through eighty on white circles centered on black squares, are encoded for compatibility with the Japanese television standard, ARIB STD B24. In that standard, they are intended to represent symbols for speed limit signs, expressed in kilometers per hour.

The Japanese era name, Reiwa (Japanese: 令和), is encoded at U+32FF SQUARE ERA NAME REIWA. The Reiwa era began on May 1, 2019. The prior era, Heisei (Japanese: 平成), began on January 8, 1989 and ended on April 30, 2019. The SQUARE ERA NAME HEISEI and three additional era names are encoded in the range U+337B..U+337E.

**CJK Compatibility: U+3300–U+33FF**

The CJK Compatibility block consists entirely of square symbols encoded for compatibility with various East Asian character sets. These come in four sets: square Latin abbreviations, Chinese telegraph symbols for hours and days, squared Katakana words, and a small set of Japanese era names.

Squared Katakana words are Katakana-spelled words that fill a single display cell (em-square) when intermixed with CJK ideographs. Likewise, the square Latin abbreviation symbols are designed to fill a single character position when mixed with CJK ideographs. Note that modern software for the East Asian market can often support the comparable functionality via styles that allow typesetting of arbitrary Katakana words or Latin abbreviations in an em-square. Such solutions are preferred when available, as they are not limited to specific lists of encoded symbols such as those in this block.

**Japanese Era Names.** The Japanese era name symbols refer to the dates given in Table 22-9.

Table 22-9. Japanese Era Names

| Code Point | Name                    | Dates                     |
|------------|-------------------------|---------------------------|
| U+32FF     | SQUARE ERA NAME REIWA   | 2019-05-01 to present day |
| U+337B     | SQUARE ERA NAME HEISEI  | 1989-01-08 to 2019-04-30  |
| U+337C     | SQUARE ERA NAME SYOUWA  | 1926-12-25 to 1989-01-07  |
| U+337D     | SQUARE ERA NAME TAISYOU | 1912-07-30 to 1926-12-24  |
| U+337E     | SQUARE ERA NAME MEIZI   | 1868-10-23 to 1912-07-29  |

**Enclosed Alphanumeric Supplement: U+1F100–U+1F1FF**

This block contains more enclosed and square symbols based on Latin letters or digits. Many are encoded for compatibility with the Japanese television standard, ARIB STD B24; others are encoded for compatibility with the Japanese cell phone carrier emoji sets.

**Regional Indicator Symbols.** A set of 26 regional indicator symbols is encoded in the range U+1F1E6..U+1F1FF. These 26 symbols correspond to a set of Latin letters A through Z, but they do not have letter properties and are not cased. They are intended for use in pairs to represent ISO 3166 region codes. This mechanism does not supplant actual ISO 3166 region codes, which simply use Latin letters from the ASCII range. Pairs of regional indica-

tor symbols should not be construed as *being* region codes (or “country codes”); rather, they constitute convenient indexes into a 26 x 26 array whose elements can be associated with region codes for the purposes of identification, processing, and rendering.

The representative glyph for a single regional indicator symbol is just a dotted box containing a capital Latin letter. The Unicode Standard does not prescribe how the pairs of regional indicator symbols should be rendered. However, current industry practice widely interprets pairs of regional indicator symbols as representing a flag associated with the corresponding ISO 3166 region code. This practice is detailed in the separate Unicode Technical Standard #51, “Unicode Emoji.” That specification includes data tables that list precisely which pairs are interpreted for any given version of UTS #51. Charts are also available showing representative flag glyphs for these interpreted pairs, displayed as part of the emoji symbol sets for many mobile platforms.

Conformance to the Unicode Standard does not require conformance to UTS #51. However, the interpretation and display of pairs of regional indicator symbols as specified in UTS #51 is now widely deployed, so in practice it is not advisable to attempt to interpret pairs of regional indicator symbols as representing anything other than an emoji flag.

Regional indicator symbols have specialized properties and behavior related to segmentation, which help to keep interpreted pairs together for line breaking, word segmentation, and so forth.

The file `EmojiSources.txt` in the Unicode Character Database provides more information about source mappings from pairs of regional indicator symbols to flag emoji in older carrier emoji sets. Provision of roundtrip mappings to those flag emoji was the original impetus to include regional indicator symbols in the Unicode Standard.

***Creative Commons License Symbols.*** Creative Commons license symbols are widely used across web platforms, content creation tools, and search engines to describe a variety of functions, permissions, and concepts related to intellectual property. The set of seven symbols was designed to work efficiently on printed pages, web pages, and signage while following the pattern of a graphic form within a circle.

Six of the seven symbols are encoded in two ranges (U+1F10D..U+1F10FF and U+1F16D..U+1F16F). One Creative Commons symbol, the circled equals sign, is represented by U+229C CIRCLED EQUALS.

### ***Enclosed Ideographic Supplement: U+1F200–U+1F2FF***

This block consists mostly of enclosed ideographic symbols. It also contains some additional squared Katakana word symbols. Most of the symbols in this block are either encoded for compatibility with the Japanese television standard ARIB STD B24, and intended primarily for use in closed captioning, or are encoded for compatibility with the Japanese cell phone carrier emoji sets.

The enclosed ideographic symbols in the range U+1F210..U+1F23B are enclosed in a square, instead of a circle. One subset of these are symbols referring to broadcast terminology, and the other subset are symbols used in baseball in Japan.

The enclosed ideographic symbols in the range U+1F240..U+1F248 are enclosed in tortoise shell brackets, and are also used in baseball scoring in Japan.

The circled ideographic symbols in the range U+1F260..U+1F265 are felicitous symbols commonly associated with Chinese folk religion. Five of these are collectively referred to as the “five-fold happiness,” representing luck, prosperity, longevity, happiness, and wealth. The sixth, U+1F264, represents “double-happiness,” a doubled variant of the happiness symbol, associated with love and marriage. Each of these symbols is paired with a respective deity in traditional folk religion.

