Chapter 24

About the Code Charts

Disclaimer
Character images shown in the code charts are not prescriptive. In actual fonts, considerable variations are to be expected.

The Unicode code charts present the characters of the Unicode Standard. This chapter explains the conventions used in the code charts and provides other useful information about the accompanying names lists.

Characters are organized into related groups called blocks (see D10b in Section 3.4, Characters and Encoding). Many scripts are fully contained within a single block, but other scripts, including some of the most widely used scripts, have characters divided across several blocks. Separate blocks contain common punctuation characters and different types of symbols.

A character names list follows the code chart for each block. The character names list itemizes every character in that block and provides supplementary information in many cases. A full list of the character names and associated annotations, formatted as a text file, NamesList.txt, is available in the Unicode Character Database. That text file contains syntax conventions which are used by the tooling that formats the PDF versions of the code charts and character names lists. For the full specification of those conventions, see NamesList.html in the Unicode Character Database.

An index to distinctive character names can also be found on the Unicode website.

For information about access to the code charts, the character name index, and the roadmap for future allocations, see Appendix B.3, Other Unicode Online Resources.
24.1 Character Names List

The following illustration exemplifies common components found in entries in the character names list. These and other components are described in more detail in the remainder of this section.

<table>
<thead>
<tr>
<th>code</th>
<th>image</th>
<th>entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>00AE</td>
<td>®</td>
<td>REGISTERED SIGN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>= registered trade mark sign (1.0)</td>
</tr>
<tr>
<td>00AF</td>
<td>-</td>
<td>MACRON</td>
</tr>
<tr>
<td></td>
<td></td>
<td>= overline, APL overbar</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• this is a spacing character</td>
</tr>
<tr>
<td></td>
<td></td>
<td>→ 02C9 &quot; modifier letter macron</td>
</tr>
<tr>
<td></td>
<td></td>
<td>→ 0304 ̂ combining macron</td>
</tr>
<tr>
<td></td>
<td></td>
<td>→ 0305 ̅ combining overline</td>
</tr>
<tr>
<td></td>
<td></td>
<td>≈ 0020 ⤝ 0304 ⃊</td>
</tr>
<tr>
<td>00E5</td>
<td>à</td>
<td>LATIN SMALL LETTER A WITH RING ABOVE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Danish, Norwegian, Swedish, Walloon</td>
</tr>
<tr>
<td></td>
<td></td>
<td>= 0061 a 030A ̂</td>
</tr>
<tr>
<td>2272</td>
<td>≤</td>
<td>LESS-THAN OR EQUIVALENT TO</td>
</tr>
<tr>
<td></td>
<td></td>
<td>~ 2272 FE00 following the slant of the lower leg</td>
</tr>
</tbody>
</table>

Images in the Code Charts and Character Lists

Each character in these code charts is shown with a representative glyph. A representative glyph is not a prescriptive form of the character, but rather one that enables recognition of the intended character to a knowledgeable user and facilitates lookup of the character in the code charts. In many cases, there are more or less well-established alternative glyphic representations for the same character.

Designers of high-quality fonts will do their own research into the preferred glyphic appearance of Unicode characters. In addition, many scripts require context-dependent glyph shaping, glyph positioning, or ligatures, none of which is shown in the code charts. The Unicode Standard contains many characters that are used in writing minority languages or that are historical characters, often used primarily in manuscripts or inscriptions. Where there is no strong tradition of printed materials, the typography of a character may not be settled. Because of these factors, the glyph image chosen as the representative glyph in these code charts should not be considered a definitive guide to best practice for typographical design.

Fonts. The representative glyphs for the Latin, Greek, and Cyrillic scripts in the code charts are based on a serifed, Times-like font. For non-European scripts, typical typefaces were selected that allow as much distinction as possible among the different characters.
The fonts used for other scripts are similar to Times in that each represents a common, widely used design, with variable stroke width and serifs or similar devices, where applicable, to show each character as distinctly as possible. Sans-serif fonts with uniform stroke width tend to have less visibly distinct characters. In the code charts, sans-serif fonts are used for archaic scripts that predate the invention of serifs, for example.

**Alternative Forms.** Some characters have alternative forms. For example, even the ASCII character U+0061 latin small letter a has two common alternative forms: the “a” used in Times and the “α” that occurs in many other font styles. In a Times-like font, the character U+03A5 greek capital letter upsilon looks like “Y”; the form Υ is common in other font styles.

A different case is U+010F latin small letter d with caron, which is commonly typeset as ð instead of ð. In such cases, the code charts show the more common variant in preference to a more didactic archetypical shape.

Many characters have been unified and have different appearances in different language contexts. The shape shown for U+2116 numero sign is a fullwidth shape as it would be used in East Asian fonts. In Cyrillic usage, № is the universally recognized glyph. See Figure 22-2.

In certain cases, characters need to be represented by more or less condensed, shifted, or distorted glyphs to make them fit the format of the code charts. For example, U+0D10 malayalam letter ai is shown in a reduced size to fit the character cell.

When characters are used in context, the surrounding text gives important clues as to identity, size, and positioning. In the code charts, these clues are absent. For example, U+2075 superscript five is shown much smaller than it would be in a Times-like text font.

Whenever a more obvious choice for representative glyph may be insufficient to aid in the proper identification of the encoded character, a more distinct variant has been selected as representative glyph instead.

**Orientation.** Representative glyphs for character in the code charts are oriented as they would normally appear in text with the exception of scripts which are predominantly laid out in vertical lines, as for Mongolian and Phags-pa. Commercial production fonts show Mongolian glyphs with their images turned 90 degrees counterclockwise, which is the appropriate orientation for Mongolian text that is laid out horizontally, such as for embedding in horizontally formatted, left-to-right Chinese text. For normal vertical display of Mongolian text, layout engines typically lay out horizontally, and then rotate the formatted text 90 degrees clockwise. Starting with Unicode 7.0, the code charts display Mongolian glyphs in their horizontal orientation, following the conventions of commercial Mongolian fonts. Glyphs in the Phags-pa code chart are treated similarly.
Special Characters and Code Points

The code charts and character lists use a number of notational conventions for the representation of special characters and code points. Some of these conventions indicate those code points which are not assigned to encoded characters, or are permanently reserved. Other conventions convey information about the type of character encoded, or provide a possible fallback rendering for non-printing characters.

Combining Characters. Combining characters are shown with a dotted circle. This dotted circle is not part of the representative glyph and it would not ordinarily be included as part of any actual glyph for that character in a font. Instead, the relative position of the dotted circle indicates an approximate location of the base character in relation to the combining mark.

093F  á  DEVANAGARI VOWEL SIGN I
• stands to the left of the consonant

0940  ə  DEVANAGARI VOWEL SIGN II

0941  ɔ  DEVANAGARI VOWEL SIGN U

The detailed rules for placement of combining characters with respect to various base characters are implemented by the selected font in conjunction with the rendering system.

During rendering, additional adjustments are necessary. Accents such as U+0302 COMBINING CIRCUMFLEX ACCENT are adjusted vertically and horizontally based on the height and width of the base character, as in “ mã ” versus “ mâ ”.

If the display of a combining mark with a dotted circle is desired, U+25CC ⊁ DOTTED CIRCLE is often chosen as the base character for the mark.

Dashed Box Convention. There are a number of characters in the Unicode Standard which in normal text rendering have no visible display, or whose only effect is to modify the display of other characters in proximity to them. Examples include space characters, control characters, and format characters.

To make such characters easily recognizable and distinguishable in the code charts and in any discussion about the characters, they are represented by a square dashed box. This box surrounds a short mnemonic abbreviation of the character’s name. For control codes which do not have a listed abbreviation to serve as a mnemonic, the representative glyph shows XXX inside the dashed box as a placeholder.

0020  sp  SPACE
• sometimes considered a control code
• other space characters: 2000 sp - 200A sp

Where such characters have a typical visual appearance in some contexts, an additional representative image may be used, either alone or with a mnemonic abbreviation.

00AD  soft  SOFT HYPHEN
= discretionary hyphen
• commonly abbreviated as SHY
This convention is also used for some graphic characters which are only distinguished by special behavior from another character of the same appearance, or which are subject to unusual rendering requirements.

2011 \[\text{non-breaking hyphen}\]  
→ 002D - hyphen-minus  
→ 00AD \[\text{soft hyphen}\]  
≈ <noBreak> 2010 -

0D4E \[\text{malayalam letter dot reph}\]  
• not used in reformed modern Malayalam orthography

The dashed box convention also applies to the glyphs of combining characters which have no visible display of their own, such as variation selectors (see Section 23.4, Variation Selectors).

FE00 \[\text{variation selector-1}\]  
• these are abbreviated VS1, and so on

Sometimes, the combining status of the character is indicated by including a dotted circle inside the dashed box, for example for \textit{viramas} that are intended to be invisible themselves, but which create the conjunct forms of adjacent consonants.

17D2 \[\text{khmer sign coeng}\]  
• functions to indicate that the following Khmer letter is to be rendered subscripted  
• shape shown is arbitrary and is not visibly rendered

Even though the presence of the dashed box in the code charts indicates that a character is likely to be a space character, a control character, a format character, or a combining character, it cannot be used to infer the actual General_Category value of that character.

\textbf{Reserved Characters.} Character codes that are marked “<reserved>” are unassigned and reserved for future encoding. Reserved codes are indicated by a \[\text{glyph}\]. To ensure readability, many instances of reserved characters have been suppressed from the names list. Reserved codes may also have cross references to assigned characters located elsewhere.

2073 \[\text{<reserved>}\]  
→ 00B3 \[^3\text{superscript three}\]

\textbf{Noncharacters.} Character codes that are marked “<not a character>” refer to noncharacters. They are designated code points that will never be assigned to a character. These codes are indicated by a \[\text{glyph}\]. Noncharacters are shown in the code charts only where they occur together with other characters in the same block. For a complete list of noncharacters, see Section 23.7, Noncharacters.

FFFF \[<not a character>\]

\textbf{Deprecated Characters.} Deprecated characters are characters whose use is strongly discouraged, but which are retained in the standard indefinitely so that existing data remain well defined and can be correctly interpreted. (See D13 in Section 3.4, Characters and Encoding.) Deprecated characters are explicitly indicated in the Unicode code charts using annotations or subheads.
**Character Names**

The character names in the code charts precisely match the normative character names in the Unicode Character Database. Character names are unique and stable. By convention, they are in uppercase. For more information on character names, see Section 4.8, Name.

**Informative Aliases**

An informative alias is an informal, alternate name for a character. Aliases are provided to assist in the correct identification of characters, in some cases providing more commonly known names than the normative character name used in the standard. For example:

```
002E .   FULL STOP
         = period, dot, decimal point
```

Informative aliases are indicated with a “=” in the names list, and by convention are shown in lowercase, except when they include a proper name. (Note that a “=” in the names list may also introduce a normative alias, which is distinguished from an informative alias by being shown in uppercase. See the following discussion of normative aliases.)

Multiple aliases for a character may be given in a single informative alias line, in which case each alias is separated by a comma. In other cases, multiple informative alias lines may appear in a single entry. Informative aliases can be used to indicate distinct functions that a character may have; this is particularly common for symbols. For example:

```
2206 ∆   INCREMENT
         = Laplace operator
         = forward difference
         = symmetric difference of sets
```

In some complex cases involving many informative aliases, rather than introduce a separate line for each set of related aliases, an informative alias line may also separate groups of aliases with semicolons:

```
1F70A ✬   ALCHEMICAL SYMBOL FOR VINEGAR
         = crucible; acid; distill; atrament; vitriol; red sulfur; borax; wine; alkali salt; mercurius vivus, quick silver
```

Informative aliases for different characters are not guaranteed to be unique. They are maintained editorially, and may be changed, added to, or even be deleted in future versions of the standard, as information accumulates about particular characters and their uses.

Informative aliases may serve as useful alternate choices for identifying characters in user interfaces. The formal character names in the standard may differ in unexpected ways from the more commonly used names for the characters. For example:

```
00B6 ¶   PILCROW SIGN
         = paragraph sign
```

**Unicode 1.0 Names.** Some character names from The Unicode Standard, Version 1.0 are indicated in the names list. These are provided only for their historical interest. Where they
About the Code Charts

24.1 Character Names List

They also are introduced with a “=”, and are shown in lowercase. In addition they are explicitly annotated with a following “1.0” in parentheses. For example:

01C3  \( \text{LATIN LETTER RETROFLEX CLICK} \)

= latin letter exclamation mark (1.0)

If a Unicode 1.0 name and one or more other informative aliases occurs in a single entry, the Unicode 1.0 name will be given first. For example:

00A6  \( \text{BROKEN BAR} \)

= broken vertical bar (1.0)
= parted rule (in typography)

Note that informative aliases other than Unicode 1.0 names may also contain clarifying annotations in parentheses.

Jamo Short Names. In the Hangul Jamo block, U+1100..U+11FF, the normative jamo short names from Jamo.txt in the UCD are displayed for convenience of reference. These are also indicated with a “=” in the names list and are shown in uppercase to imply their normative status. For example:

1101  \( \text{HANGUL CHOSEONG SSANGKIYEOK} \)

= GG

The Jamo short names do not actually have the status of alternate names; instead they are simply string values associated with the jamo characters, for use by the Unicode Hangul Syllable Name Generation algorithm. See Section 3.12, Conjoining Jamo Behavior.

Normative Aliases

A normative character name alias is a formal, unique, and stable alternate name for a character. In limited circumstances, characters are given normative character name aliases where there is a defect in the character name. These normative aliases do not replace the character name, but rather allow users to refer formally to the character without requiring the use of a defective name. For more information, see Section 4.8, Name.

Normative aliases which provide information about corrections to defective character names or which provide alternate names in wide use for a Unicode format character are printed in the character names list, preceded by a special symbol \( \star \).

FE18  \( \text{PRESENTATION FORM FOR VERTICAL RIGHT WHITE LENTICULAR BRACKET} \)

\( \star \) presentation form for vertical right white lenticular bracket
= misspelling of “BRACKET” in character name is a known defect
≈ <vertical> 3017

Normative aliases serving other purposes, if listed, are shown by convention in all caps, following an “=”. In contrast, informative aliases are shown in lowercase. Normative aliases of type “control” typically represent names of control functions as listed in the latest edition of ISO 6429. Normative aliases of type “figment” for control codes are not listed. Normative aliases which represent commonly used abbreviations for control codes or format characters are shown in all caps, enclosed in parentheses. For editorial presentation in the
names list, those parenthetical listings may occur on the same lines as informative aliases. See NameAliases.txt in the UCD for the definitive listing of all normative aliases, also including their types, suitable for machine parsing.

**Cross References**

Cross references (preceded by →) are used to indicate a related character of interest, but without indicating the exact nature of the relation. Cross references are most commonly used to indicate a different character of similar or occasionally identical appearance, which might be confused with the character in question. Cross references are also used to indicate characters with similar names or functions, but with distinct appearances. Cross references may also be used to show linguistic relationships, such as letters used for transliteration in a different script. Some blocks start with a list of cross references that simply point to related characters of interest in other blocks. Examples of various types of cross references follow.

**Explicit Inequality.** The cross reference indicates that two (or more) characters are not identical, although the representative glyphs that depict them are identical or very close in appearance.

003A : colon
• also used to denote division or scale; for that mathematical use 2236 is preferred
→ 0589 : armenian full stop
→ 05C3 : hebrew punctuation sof pasuq
→ 2236 : ratio
→ A789 : modifier letter colon

**Related Functions.** The cross reference indicates that two (or more) characters have similar functions, although the representative glyphs are distinct. See, for example, the cross references to division slash, divides, and ratio in the names list entry for U+00F7 division sign:

00F7 ÷ division sign
= obelus
• occasionally used as an alternate, more visually distinct version of 2212 or 2011 in some contexts
• historically used as a punctuation mark to denote questionable passages in manuscripts
→ 070B – syriac harklean obelus
→ 2052 % commercial minus sign
→ 2212 – minus sign
→ 2215 / division slash
→ 2223 \ divides
→ 2236 : ratio
→ 2797 ⊳ heavy division sign

In addition to related mathematical functions, cross references may show other related functions, such as use of distinct symbols in different phonetic transcription systems to represent the same sounds. For example, the cross reference to U+0296 in the following entry shows the IPA equivalent for U+01C1:
01C1 | LATIN LETTER LATERAL CLICK
     = double pipe
     • Khoisan tradition
     • “x” in Zulu orthography
     → 0296 ↓ latin letter inverted glottal stop
     → 2225 ℏ parallel to

Related Names. The cross reference indicates that two (or more) characters have similar
and possibly confusable names, although their appearance is distinct.

1F32B ⇒ FOG
     → 1F301 ☇ foggy

Transliteration. The cross reference indicates a character from another script commonly
used for transliteration of the character in question. Note that this use of cross references is
deliberately limited to a few special cases such as Mongolian:

182E ⸜ MONGOLIAN LETTER MA
     → 043C O cyrillic small letter em

This use of cross references is also seen for compatibility digraph letters for Serbo-Croatian:

01C9 lj LATIN SMALL LETTER LJ
     → 0459 ç cyrillic small letter lje

Blind Cross References. The cross reference notation is also used to point to related char-
acters in other blocks. In these cases, the cross reference is not from any particular code
point. For example, the list of cross references at the top of the Currency Symbols block
points to many other currency signs scattered throughout the standard.

In a few instances, a cross reference points from a reserved, unassigned code point. These
cross references occur in cases where the structure of a chart might lead a user to expect a
particular character at a code point, but the character to use is actually encoded elsewhere.
This occurs, for example, in several Indic blocks to point to the shared danda characters:

For viram punctuation, use the generic Indic 0964 and 0965.

0A64 <reserved>
     → 0964 l devanagari danda

0A65 <reserved>
     → 0965 ℂ devanagari double danda

Cross references are neither exhaustive nor symmetric. Typically a general character would
have cross references to more specialized characters, but not the other way around.

Information About Languages

An informative note may include a list of one or more of the languages using that character
where this information is considered useful. For case pairs, the annotation is given only for
the lowercase form to avoid needless repetition. An ellipsis “...” indicates that the listed lan-
guages cited are merely the principal ones among many.
Case Mappings

When a case mapping corresponds solely to a difference based on small versus capital in the names of the characters, the case mapping is not given in the names list but only in the Unicode Character Database.

0041 A LATIN CAPITAL LETTER A
01F2 Dz LATIN CAPITAL LETTER D WITH SMALL LETTER Z ∼ 0044 D 007A z

When the case mapping cannot be predicted from the name, the casing information is sometimes given in a note.

00DF ß LATIN SMALL LETTER SHARP S
  = Eszett
  • German
  • not used in Swiss High German
  • uppercase is “SS” or ß 1E9E
  • typographically the glyph for this character can be based on a ligature of 017F Ï with either 0073 s or with an old-style glyph for 007A z (the latter similar in appearance to 3 0292). Both forms exist interchangeably today.
  → 03B2 β greek small letter beta

For more information about case and case mappings, see Section 4.2, Case.

Decompositions

The decomposition sequence (one or more letters) given for a character is either its canonical mapping or its compatibility mapping. The canonical mapping is marked with an identical to symbol ≡.

00E5 à LATIN SMALL LETTER A WITH RING ABOVE
  • Danish, Norwegian, Swedish, Walloon
  ≡ 0061 a 030A 

212B Å ANGSTROM SIGN
  ≡ 00C5 Ä latin capital letter a with ring above

Compatibility mappings are marked with an almost equal to symbol ∼. Formatting information may be indicated with a formatting tag, shown inside angle brackets.

01F2 Dz LATIN CAPITAL LETTER D WITH SMALL LETTER Z ∼ 0044 D 007A z
FF21 A FULLWIDTH LATIN CAPITAL LETTER A ∼ <wide> 0041 A
The following compatibility formatting tags are used in the Unicode Character Database:

- `<font>` A font variant (for example, a blackletter form)
- `<noBreak>` A no-break version of a space, hyphen, or other punctuation
- `<initial>` An initial presentation form (Arabic)
- `<medial>` A medial presentation form (Arabic)
- `<final>` A final presentation form (Arabic)
- `<isolated>` An isolated presentation form (Arabic)
- `<circle>` An encircled form
- `<super>` A superscript form
- `<sub>` A subscript form
- `<vertical>` A vertical layout presentation form
- `<wide>` A fullwidth (or zenkaku) compatibility character
- `<narrow>` A halfwidth (or hankaku) compatibility character
- `<small>` A small variant form (CNS compatibility)
- `<square>` A CJK squared font variant
- `<fraction>` A vulgar fraction form
- `<compat>` Otherwise unspecified compatibility character

In the character names list accompanying the code charts, the “<compat>” label is suppressed, but all other compatibility formatting tags are explicitly listed in the compatibility mapping.

Decomposition mappings are not necessarily full decompositions. For example, the decomposition for U+212B Â Ångström sign can be further decomposed using the canonical mapping for U+00C5 Å Latin capital letter a with ring above. (For more information on decomposition, see Section 3.7, Decomposition.)

Compatibility decompositions do not attempt to retain or emulate the formatting of the original character. For example, compatibility decompositions with the `<noBreak>` formatting tag do not use U+2060 WORD JOINER to emulate nonbreaking behavior; compatibility decompositions with the `<circle>` formatting tag do not use U+20DD COMBINING ENCLOSING CIRCLE; and compatibility decompositions with formatting tags `<initial>`, `<medial>`, `<final>`, or `<isolater` for explicit positional forms do not use ZWJ or ZWNJ. The one exception is the use of U+2044 FRACTION SLASH to express the `<fraction>` semantics of compatibility decompositions for vulgar fractions.

**Standardized Variation Sequences**

The Unicode Standard defines a number of standardized variation sequences. These consist of a single base character followed by a variation selector. Use of a standardized variation sequence allows a user to indicate their preference for a display with a particular glyph or subset of glyphs for the given character.
In the character names list, each variation sequence for standardized variants is listed in
the entry for the base character for that sequence. In some cases a character may be associ-
ated with multiple variation sequences. A standardized variation sequence is identified in
the character names list with an initial swung dash “~”.

\[
\begin{align*}
228A & \subset \text{ SUBSET OF WITH NOT EQUAL TO} \\
& \sim 228A \text{ FE00 } \subset \text{ with stroke through bottom members}
\end{align*}
\]

The glyphs for most emoji variation sequences cannot be displayed by the font technology
used to print the code charts. In those cases, the glyphs for the standardized variation
sequences are omitted from the names list. Representative glyphs for both the colorful
emoji presentation style and the text style of all emoji variation sequences can be found in
the emoji charts section of the Unicode website.

Characters for which one or more standardized variants have been defined are displayed in
the code charts with a special convention: the code chart cell for such characters has a
small black triangle in its upper-right corner.

![Character with black triangle](image)

Characters which have one or more positional glyph variants, but no standardized variants
have a small open triangle in the upper-right corner of their code chart cell.

Blocks containing characters for which standardized variation sequences and/or positional
glyph variants are shown in the names list also have a separate summary listing at the end
of the block, displaying the variants in a large font size. Each entry in these summary list-
ings is shown as follows:

\[
\begin{align*}
228A & \subset \text{ SUBSET OF WITH NOT EQUAL TO} \\
& \sim 228A \text{ FE00 } \subset \text{ with stroke through bottom members}
\end{align*}
\]

The list of standardized variation sequences in the character names list matches the list
defined in the data file StandardizedVariants.txt in the Unicode Character Database. Emoji
variation sequences are not included in these summary listings at the ends of blocks,
because of the limitations in font technology used for the code chart display. Ideographic
variation sequences defined in the Ideographic Variation Database are also not included.
See Section 23.4, Variation Selectors for more information.

Standardized Variation Sequences to select glyphs appropriate for display of CJK compati-
bility ideographs are shown not with the corresponding CJK unified ideograph, but rather
with the CJK compatibility ideograph defining the glyph to be selected. All CJK compatibil-
ity ideographs have a canonical decomposition to a CJK unified ideograph for historical
reasons. This means that direct use of CJK compatibility ideographs is problematical, because they are not stable under normalization. To indicate that one of the compatibility glyph shapes is desired, the indicated variation selector can be used with the CJK unified ideograph. In the CJK Compatibility Ideographs and CJK Compatibility Supplement blocks, the canonical decomposition and the relevant standardized variation sequence are shown together with respective representative glyphs for the sources defined for the CJK compatibility ideograph; see Figure 24-3.

Note that there are no indications of variation sequences in the charts for CJK unified ideographs. See the Ideographic Variation Database (IVD) for information on registered variation sequences for CJK unified ideographs.

**Positional Forms**

In cursive scripts which have contextually defined positional forms for letters, such as Arabic or Mongolian, the basic positional forms may appear in the code charts. Such forms, when they occur, appear in the charts in the summary listings, together with any standardized variation sequences. In Versions 9.0 through 12.1, such positional forms were included in the code chart for Mongolian, but have been removed from the code charts starting with Version 13.0, with the intent that they be shown instead in a publication dedicated to the details of the Mongolian text model.

**Block Headers**

The code charts are segmented by the format tooling into blocks. (See Definition D10b in Section 3.4, Characters and Encoding.) The page headers for the code charts are based on the normative values of the Block property defined in Blocks.txt in the Unicode Character Database, with a few exceptions. For example, the ASCII and Latin-1 ranges have their block headers adjusted editorially to reflect the presence of C0 and C1 control characters in those ranges. This means that the Block property value for the block associated with the range U+0080..U+00FF is "Latin-1 Supplement", but the block header used in the code charts is "C1 Controls and Latin-1 Supplement".

The start and end code points printed in the block headers in the code charts and character names list reflect the ranges that are printed on that page, and thus should not be confused with the normative ranges listed in Blocks.txt.

On occasion, the code chart format tooling also introduces artificial block headers to enable the display of code charts for noncharacters that are outside the range of any normative block range. For example, the two noncharacters U+3FFFE..U+3FFFF are artificially displayed in a code chart with a block header "Unassigned", showing a range U+3FF80..U+3FFFF.

As a result of these and other editorial considerations, implementers are cautioned not to attempt to pull block range values from the code charts, nor to attempt to parse them from the NamesList.txt file in the Unicode Character Database. Instead, normative values for block ranges and names should always depend on Blocks.txt.
Subheads

The character names list contains a number of informative subheads that help divide up the list into smaller sublists of similar characters. For example, in the Miscellaneous Symbols block, U+2600..U+26FF, there are subheads for "Astrological symbols," "Chess symbols," and so on. Such subheads are editorial and informative; they should not be taken as providing any definitive, normative status information about characters in the sublists they mark or about any constraints on what characters could be encoded in the future at reserved code points within their ranges. The subheads are subject to change.
24.2 CJK and Other Ideographs

The code charts for CJK and Tangut ideographs differ significantly from those for other characters in the standard.

**CJK Unified Ideographs**

Character names are not provided for any of the code charts of CJK Unified Ideograph character blocks, because the name of a unified ideograph simply consists of its Unicode code point preceded by `cjk unified ideograph-`.

In other code charts, each character is shown with a single representative glyph, but in the code charts for CJK Unified and Compatibility Ideographs, each character may have multiple representative glyphs. Each character is shown with as many representative glyphs as there are Ideographic Research Group (IRG) sources defined for that character. The representative glyph for each IRG source is not necessarily the only preferred glyph for the corresponding region, and developers are therefore encouraged to refer to regional standards or typographical conventions to determine the appropriate glyph. Each representative glyph is accompanied with its source reference provided in alphanumeric form. Altogether, there are eleven IRG sources, as shown in Table 24-1. Data for these IRG sources are documented in Unicode Standard Annex #38, “Unicode Han Database (Unihan).”

<table>
<thead>
<tr>
<th>Name</th>
<th>Source Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>G source</td>
<td>China PRC and Singapore</td>
</tr>
<tr>
<td>H source</td>
<td>Hong Kong SAR</td>
</tr>
<tr>
<td>J source</td>
<td>Japan</td>
</tr>
<tr>
<td>KP source</td>
<td>North Korea</td>
</tr>
<tr>
<td>K source</td>
<td>South Korea</td>
</tr>
<tr>
<td>M source</td>
<td>Macao SAR</td>
</tr>
<tr>
<td>S source</td>
<td>SAT</td>
</tr>
<tr>
<td>T source</td>
<td>TCA</td>
</tr>
<tr>
<td>UK source</td>
<td>UK</td>
</tr>
<tr>
<td>U source</td>
<td>Unicode</td>
</tr>
<tr>
<td>V source</td>
<td>Vietnam</td>
</tr>
</tbody>
</table>

To assist in reference and lookup, each CJK Unified Ideograph is accompanied by a representative glyph of its Unicode radical and by its Unicode radical-stroke counts. These are printed directly underneath the Unicode code point for the character. A radical-stroke index to all of the CJK ideographs is also provided separately on the Unicode website.

**Chart for the Main CJK Block.** For the CJK Unified Ideographs block (U+4E00..U+9FFF) the glyphs are arranged in the following order: G, H, and T sources are grouped under the header “C.” J, K, and V sources are listed under their respective headers. Each row contains positions for all six sources, and if a particular source is undefined for CJK Unified Ideo-
graphs, that position is left blank in the row. This format is illustrated by Figure 24-1. In general, if a character has an M, U, UK or S source but none of the other kinds of sources, the source is shown in the G, H, T or J source position respectively. If any of the remaining sources are present, the M, U, UK or S sources are shown on a line by themselves below the G, H, T or J source position, respectively. Note that this block does not currently contain any characters with UK or S sources. The KP sources are not shown due to lack of reliable glyph information.

**Figure 24-1. CJK Chart Format for the Main CJK Block**

<table>
<thead>
<tr>
<th>HEX</th>
<th>C</th>
<th>J</th>
<th>K</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>4F1A</td>
<td>人</td>
<td>94</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G0-3B61</td>
<td>H-864E</td>
<td>T3-2275</td>
<td>J0-3271</td>
</tr>
</tbody>
</table>

**Charts for CJK Extensions.** The code charts for all of the extension blocks for CJK Unified Ideographs use a more condensed format. That format dispenses with the “C, J, K, V, and H” headers and leaves no holes for undefined sources. For those blocks, sources are always shown in the following order: G, T, J, K, KP, V, H, M, U, UK, and S. The first letters of the source information provide the source type for all sources except G. KP sources are omitted from the code charts because of the lack of an appropriately vetted font for display.

The multicolumn code charts for CJK Extensions A and B use the condensed format with three source columns per entry, and with entries arranged in three columns per page. An entry may have additional rows, if it is associated with more than three sources, as illustrated in Figure 24-2 for CJK Extension A.

**Figure 24-2. CJK Chart Format for CJK Extension A**

The multicolumn code charts for the CJK Unified Ideographs Extension B block (U+20000..U+2A6DF), which are formatted like those for the Extension A block, were introduced in Version 5.2 of the standard. From Version 6.1 through 13.0 of the standard, those multicolumn code charts had the additional idiosyncrasy that the first source shown always corresponded to the “UCS2003” representative glyph. Those representative glyphs were the only ones used up through Version 5.1 of the standard for that block, and have been archived as separate code charts with a single representative glyph.

The multicolumn code charts for the other extension blocks for CJK Unified Ideographs use the condensed format with two source columns per entry, and with entries arranged in
four columns per page. An entry may have additional rows if it is associated with more than two sources.

Compatibility Ideographs

The format of the code charts for the CJK Compatibility Ideograph blocks is largely similar to the CJK chart format for Extension A, as illustrated in Figure 24-3. However, several additional notational elements described in Section 24.1, Character Names List are used. In particular, for each CJK compatibility ideograph other than the small list of unified ideographs included in these charts, a canonical decomposition is shown. The ideographic variation sequence for each compatibility CJK ideograph is listed below the canonical decomposition, introduced with a tilde sign.

Figure 24-3. CJK Chart Format for Compatibility Ideographs

| FA15 | 齥 齥 |
| 15.14 | 3B79A | UTC-00918 |
| = 51DE | 齥 |
| ~ 51DE FE00 |

The twelve CJK unified ideographs in the CJK Compatibility Ideographs block have no canonical decompositions or corresponding ideographic variation sequences; instead, each is clearly labeled with an annotation identifying it as a CJK unified ideograph.

Figure 24-4. Annotations Identifying CJK Unified Ideographs

| FA24 | 迺 迺 迺 |
| 12.24 | 3B79F | UTC-00851 |
| = a CJK unified ideograph |

Character names are not provided for any CJK Compatibility Ideograph blocks because the name of a compatibility ideograph simply consists of its Unicode code point preceded by CJK COMPATIBILITY IDEOGRAPH-.

Tangut Ideographs

Code charts for Tangut ideographs use the same condensed format as the code charts for CJK Extension A, but with a single source column per entry, and with entries arranged in five columns per page.
Character names are not provided for any of the code charts of Tangut character blocks; the name of each Tangut ideograph simply consists of its Unicode code point preceded by \texttt{TANGUT IDEOGRAPH-}. 
24.3 Hangul Syllables

As in the case of CJK Unified Ideographs, a character names list is not provided for the online chart of characters in the Hangul Syllables block, U+AC00..U+D7AF, because the name of a Hangul syllable can be determined by algorithm as described in Section 3.12, Conjoining Jamo Behavior. The short names used in that algorithm are listed in the code charts as aliases in the Hangul Jamo block, U+1100..U+11FF, as well as in Jamo.txt in the Unicode Character Database.
About the Code Charts

966

24.3  Hangul Syllables