About This Book

This book is the authoritative source of information on the Unicode character encoding standard (henceforth referred to simply as "the Unicode standard"), an international character code for information processing. The aim of the creators of the Unicode standard is to encode all characters used for written communication, both modern and historic, in a simple and consistent manner. Version 1.0 of the Unicode standard includes all major scripts used in the world; future editions will add less commonly used scripts and those of primarily historical interest.

The Unicode standard had its genesis in early 1988 when a group of information professionals with extensive experience in multilingual computing agreed that no encoding methodology used in their field possessed the elegance and simplicity of ASCII. The Unicode character encoding was established as a fixed-width encoding of 16 bits, which would provide a sufficient number of unique codes for the world's scripts and technical symbols in common use, and at the same time promote efficient and flexible system design.

In January 1991, The Unicode Consortium was incorporated as *Unicode, Inc.*, a non-profit organization whose charter is to maintain and promote the Unicode standard worldwide. Founding members of the Consortium include major companies and institutions involved in international computing. Membership is open to any organization interested in contributing to the design, implementation, and maintenance of the Unicode standard.

How to Use This Book

The Unicode Standard: Worldwide Character Encoding, Version 1.0 is divided into two volumes. Volume One contains introductory material, coding architecture, and conformance requirements, as well as code charts and names lists for all non-ideographic characters and a large number of supplementary mapping tables for the non-ideographic characters.

Volume Two is devoted to the East Asian (Han) ideographic characters. It contains code charts and cross-tabulations with equivalent characters from existing national and corporate standards. It also contains a radical/stroke index to the ideographic characters.

The essential component of Volume One is the coding charts. The complete Unicode character repertoire is laid out in successive panels of 256 code points. For the convenience of users of the Unicode character encoding standard, codes have been grouped by linguistic and functional categories into character blocks. Section 3.1, General Scripts, describes the scope of each block and the origin of its characters. In the section following the code charts (Section 3.8), all non-ideographic characters, with the exception of Korean Hangul syllables, are named.

The principles and architecture of the Unicode standard are covered in the initial chapters that precede the coding charts. A complete explanation of the design and structure of the Unicode standard is provided in Section 2.2 of this book. This chapter also includes information on significant implementation issues, such as directionality, which affect conformance of an implementation. A separate appendix, Implementation Guidelines, addresses common questions.

The coding charts alone are not sufficient to implement the Unicode character encoding scheme. Characters must be categorized according to specific properties, such as case mapping, or directional properties which can be found in Chapter 4, Character Properties.

Additional tables are provided in Chapter 5 to promote uniform mapping in specific circumstances. Tables for composed characters show how text elements may be made up of sequences of Unicode characters.

To facilitate the introduction and widespread adoption of the Unicode standard, mapping tables are provided in Chapter 6 to link characters from the Unicode standard to characters in other standards; specifically, ISO standards 8859, 8879 (SGML), and DIS 6862.2, JIS, and some vendor encodings. Additional mapping tables will be included in future editions as needed.

There are two indexes at the back of the book: an index to non-ideographic character names and a general index. The index to character names includes alternative names (aliases) in addition to the formal Unicode standard designation for each character. The general index covers broader topics,

and gives the page where information on the topic appears. A quick guide to all the character blocks included in version 1.0 of the Unicode standard and the code range for each can be found at the end of the book (pp. 681–82).

Please contact The Unicode Consortium if information is missing or you have difficulty using this book. The Appendix includes instructions on how to propose a new character for inclusion in the Unicode standard. You do not have to belong to The Unicode Consortium to suggest improvements to the Unicode standard.

Notations and Conventions

Codes

An individual Unicode value is expressed as U+nnnn, where nnnn is a four digit number in hexadecimal notation, using the digits 0–9 and the letters A–F (for 10 through 15 respectively).

U+0416 is the Unicode value for the character named CYRILLIC CAPITAL LETTER ZHE

A range of Unicode values is expressed as U+xxxx→U+yyyy, where xxxx and yyyy are the first and last Unicode values in the range.

U+0900 → U+097F

In charts and tables the U+ is omitted.

A sequence of Unicode values or names is expressed using a special plus sign between successive items.

U+0259 + U+02DE and SCHWA + RHOTIC HOOK

The codespace is presented as successive panels of 256 code values. The upper left and right hand corners of each chart page show the beginning and ending Unicode values.

Images

The image shown in a grid cell of a panel should not be considered to be the prescriptive form of a character, but is merely intended as a typical representation of the character encoded by the value.

U+0061 LATIN SMALL LETTER A can be represented by a or *a*.

Where a character is commonly represented in more than one way, alternate images are separated by a vertical line.

U+0024 DOLLAR SIGN $ | $ | $

A character that is shown with a dashed circle must be rendered in relation to the previous characters in the data stream.

U+0308 NON-SPACING DIAERESIS and U+093F DEVANAGARI VOWEL SIGN I

A character that is shown as text surrounded by a dashed box has no visible manifestation on its own.

U+200A HAIR SPACE

The *geta* 〓 (a missing glyph symbol) has been placed at code points in the Han character block (U+4000→) when the appropriate glyph was unavailable. (These missing glyphs will be added in a future version of the Unicode Standard.)

Names

All characters included in the Unicode standard, except for Han ideographs, have unique names. The Unicode standard names follow the character naming conventions of the International Organization for Standardization; that is, names are written only in uppercase letters of the English alphabet plus the hyphen.

Wherever possible, names are taken from published standards, or, in the absence of a published standard, follow the recommendation of an authoritative organization. Where a systematic list of names does not exist, names in the Unicode standard describe the glyphs that depict the characters.

In running text, a formal Unicode name is shown in small capitals.

GREEK SMALL LETTER MU

Alternative names for Unicode characters (aliases) appear in italics in running text,

umlaut

in italics, preceded by an equals sign, in the character lists by block,

= *stress mark*

and in lower case (no capitalization) in the Unicode index to character names.

st. andrews cross

Italics are also used in running text to refer to characters collectively,

. . . variant forms of *cedilla* . . .

to refer to a text element that is not explicitly encoded,

. . . *pasekh alef* can be composed from other characters

or to set off a foreign word.

the Welsh word *ynghyd*

Names List Annotations

The following symbols are used in the Character Names List of each character block:

= Identifies alternative names by which the character is known, that is, the names that are synonyms or aliases for the Unicode name

```
0023 # NUMBER SIGN
        = pound sign
```

x Indicates a cross-reference to another Unicode character, which is *not* synonymous.

→ Within a cross-reference, points to the Unicode value of the character referenced.

The cross-references fall into the following classes:

Explicit inequality, in which the two characters do not have the same semantic meaning although the glyphs that depict them are identical or very close.

```
003A : COLON
        x (ratio → 2236)
```

Case form mapping, in which the other case form is encoded in another character block.

```
00FF ÿ LATIN SMALL LETTER Y DIAERESIS
        x (latin capital letter y diaeresis→ 0178)
```

Other linguistic relationship

```
01C9 lj LATIN SMALL LETTER L J
        x (cyrillic small letter lje → 0459)
        The Serbo-Croatian language may be written as Croatian using Latin
        script or as Serbian using Cyrillic script.
```

The character at this position in the source standard was not included in the Unicode standard at the corresponding position; the Unicode value given in the cross-reference should be used instead.

```
0373    x (pound sign → 00A3)
```

In the Character Names List, the language(s) using a given character are noted in cases where this information may be helpful. (Such annotation is given only after the lowercase form of case pairs, to avoid needless repetition.) An ellipsis "..." indicates that the listed languages cited are merely the principal ones among many. If the annotation *does not* end with an ellipsis, then the cited list is thought to be complete.

Notice

The Unicode Standard Version 1.0 is intended to be a complete, implementable character encoding which reflects the final decisions made by the Unicode Technical Committee. However, recent developments in the international standards community have raised the possibility that by allowing some flexibility in the content of Unicode 1.0, it will be possible to reach the goal of a single, universal, international character encoding standard.

The International Standards Organization committee responsible for multibyte character encoding (ISO JTC1/SC2/WG2) has been developing a 32-bit character encoding standard known as ISO DIS 10646. During the first half of 1991, ISO member countries had voted on whether to approve 10646 as an international standard; it failed by a significant margin. A majority of the negative votes reflected a strong desire to merge 10646 and the Unicode standard into a single international standard. Recently, members of the Unicode Consortium and the ISO committee have been making considerable progress towards a merger of the two encoding standards; the Unicode Technical Committee and the officers of the Unicode Consortium also strongly support these efforts.

At the August 1991 meeting of WG2 in Geneva resolutions constituting a merger were approved. The proposed framework for the actual encoding provides for the Unicode standard as a two-byte subset of a canonical 4-byte international standard character encoding.

The international standard resulting from this cooperation between the two groups may differ in some respects from version 1.0 of the Unicode standard. Insofar as minor changes to the Unicode standard may be necessary in order to produce the merged standard, the last article of section 2.3 ("Existing characters will not be reassigned or removed") is amended to read: "Except as required for merger with ISO 10646, future versions of the Unicode standard will not reassign, rename or remove Unicode 1.0 characters."

Specifically, the following changes were agreed upon at the August WG2 meeting (too late to be reflected in this volume):

- Expanded Compatibility Zone. U+F800 → U+FDFF will be removed from the Private Use Area and be encoded as additional characters in the Compatibility Zone, in order to accomodate additional repertoire merged in from ISO DIS 10646. Implementations of Unicode 1.0 should treat this range as reserved, unassigned characters.
- The Corporate Use Zone of the Private Use Area should start at U+F7FF and utilize descending code values, rather than starting at U+FDFF.

- The unified Han ideographic characters may not start at U+4000, but at some higher value. Software should not rely on 0x4000 as a valid indicator of the edge of that zone.
- Pending reconfirmation of the encoding of ISO 10646 APL characters, use of U+2300 APL COMPOSE OPERATOR and U+2301 APL OUT is discouraged, as those characters may be withdrawn to reach merger with 10646.
- The official IS 10646 may have character names which differ from Unicode 1.0 character names in some respects.

The nature of the international standards process does not permit the schedule for the merger to be determined with precision; however, if all goes well, a new Draft International Standard should be issued by the end of 1991, and that should become a final International Standard by mid-1992. The Unicode Consortium is aware of the time pressures facing current implementers of the Unicode standard, but due to the overall benefit of merging the two encodings, it is felt necessary to maintain some flexibility in allowing the reassignment of code points for the purposes of the merger.

In the event that minor changes are made to Unicode 1.0 to accommodate merger with ISO 10646, the Unicode Consortium will designate the modified Unicode standard as the Unicode Standard, Version 1.1. (This is different than the expanded edition that has been referred to as "Version 1.1" in previous contexts.)

August 30, 1991

The officers of the Unicode Consortium:

Mark Davis, President
Mike Kernaghan, Vice President
Joe Becker, Technical Vice President
Bill English, Treasurer
Ken Whistler, Secretary
Lee Collins, Technical Director
Asmus Freytag, Technical Director