Electronic Edition

This file is part of the electronic edition of *The Unicode Standard, Version 5.0*, provided for online access, content searching, and accessibility. It may not be printed. Bookmarks linking to specific chapters or sections of the whole Unicode Standard are available at

Purchasing the Book

For convenient access to the full text of the standard as a useful reference book, we recommend purchasing the printed version. The book is available from the Unicode Consortium, the publisher, and booksellers. Purchase of the standard in book format contributes to the ongoing work of the Unicode Consortium. Details about the book publication and ordering information may be found at

http://www.unicode.org/book/aboutbook.html

Joining Unicode

You or your organization may benefit by joining the Unicode Consortium: for more information, see Joining the Unicode Consortium at

http://www.unicode.org/consortium/join.html

This PDF file is an excerpt from *The Unicode Standard, Version 5.0*, issued by the Unicode Consortiumand published by Addison-Wesley. The material has been modified slightly for this electronic editon, however, the PDF files have not been modified to reflect the corrections found on the Updates and Errata page (http://www.unicode.org/errata/). For information on more recent versions of the standard, see http://www.unicode.org/versions/enumeratedversions.html.

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.

The Unicode® Consortium is a registered trademark, and Unicode™ is a trademark of Unicode, Inc. The Unicode logo is a trademark of Unicode, Inc., and may be registered in some jurisdictions.

The authors and publisher have taken care in the preparation of this book, 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. *Dai Kan-Wa Jiten*, used as the source of reference Kanji codes, was written by Tetsuji Morohashi and published by Taishukan Shoten.

Cover and CD-ROM label design: Steve Mehallo, www.mehallo.com

The publisher offers excellent discounts on this book when ordered in quantity for bulk purchases or special sales, which may include electronic versions and/or custom covers and content particular to your business, training goals, marketing focus, and branding interests. For more information, please contact U.S. Corporate and Government Sales, (800) 382-3419, corpsales@pearsontechgroup.com. For sales outside the United States please contact International Sales, international@pearsoned.com

Visit us on the Web: www.awprofessional.com

Chapter 1

Introduction

The Unicode Standard is the universal character encoding standard for written characters and text. It defines a consistent way of encoding multilingual text that enables the exchange of text data internationally and creates the foundation for global software. As the default encoding of HTML and XML, the Unicode Standard provides a sound underpinning for the World Wide Web and new methods of business in a networked world. Required in new Internet protocols and implemented in all modern operating systems and computer languages such as Java and C#, Unicode is the basis of software that must function all around the world.

With Unicode, the information technology industry has replaced proliferating character sets with data stability, global interoperability and data interchange, simplified software, and reduced development costs.

While taking the ASCII character set as its starting point, the Unicode Standard goes far beyond ASCII's limited ability to encode only the upper- and lowercase letters A through Z. It provides the capacity to encode all characters used for the written languages of the world—more than 1 million characters can be encoded. No escape sequence or control code is required to specify any character in any language. The Unicode character encoding treats alphabetic characters, ideographic characters, and symbols equivalently, which means they can be used in any mixture and with equal facility (see *Figure 1-1*).

The Unicode Standard specifies a numeric value (code point) and a name for each of its characters. In this respect, it is similar to other character encoding standards from ASCII onward. In addition to character codes and names, other information is crucial to ensure legible text: a character's case, directionality, and alphabetic properties must be well defined. The Unicode Standard defines these and other semantic values, and it includes application data such as case mapping tables and character property tables as part of the Unicode Character Database. Character properties define a character's identity and behavior; they ensure consistency in the processing and interchange of Unicode data. See *Section 4.1, Unicode Character Database.*

Unicode characters are represented in one of three encoding forms: a 32-bit form (UTF-32), a 16-bit form (UTF-16), and an 8-bit form (UTF-8). The 8-bit, byte-oriented form, UTF-8, has been designed for ease of use with existing ASCII-based systems.

The Unicode Standard, Version 5.0, is code-for-code identical with International Standard ISO/IEC 10646. Any implementation that is conformant to Unicode is therefore conformant to ISO/IEC 10646.

Figure 1-1. Wide ASCII

ASCII/8859-1 Text			Unicode Text	
A	0100 0001		A	0000 0000 0100 0001
S	0101 0011		S	0000 0000 0101 0011
C	0100 0011		C	0000 0000 0100 0011
I	0100 1001		I	0000 0000 0100 1001
I	0100 1001		I	0000 0000 0100 1001
/	0010 1111			0000 0000 0010 0000
8	0011 1000		天	0101 1001 0010 1001
8	0011 1000		地	0101 0111 0011 0000
5	0011 0101			0000 0000 0010 0000
9	0011 1001		س	0000 0110 0011 0011
-	0010 1101		ل	0000 0110 0100 0100
1	0011 0001		ا	0000 0110 0010 0111
	0010 0000		م	0000 0110 0100 0101
t	0111 0100			0000 0000 0010 0000
e	0110 0101		α	0000 0011 1011 0001
x	0111 1000		♯	0010 0010 0111 0000
t	0111 0100		γ	0000 0011 1011 0011

The Unicode Standard contains 1,114,112 code points, most of which are available for encoding of characters. The majority of the common characters used in the major languages of the world are encoded in the first 65,536 code points, also known as the Basic Multilingual Plane (BMP). The overall capacity for more than 1 million characters is more than sufficient for all known character encoding requirements, including full coverage of all minority and historic scripts of the world.

1.1 Coverage

The Unicode Standard, Version 5.0, contains 99,024 characters from the world's scripts. These characters are more than sufficient not only for modern communication in most languages, but also for the classical forms of many languages. The standard includes the European alphabetic scripts, Middle Eastern right-to-left scripts, and scripts of Asia, as well as many others. The unified Han subset contains 70,229 ideographic characters defined by national and industry standards of China, Japan, Korea, Taiwan, Vietnam, and Singapore.

In addition, the Unicode Standard includes punctuation marks, mathematical symbols, technical symbols, geometric shapes, and dingbats.

Many new scripts have been added between Version 4.0 and Version 5.0, including N'Ko, New Tai Lue, Buginese, Glagolitic, Coptic, Tifinagh, Syloti Nagri, Balinese, Phags-pa, Old Persian, Phoenician, Kharoshthi, and Sumero-Akkadian Cuneiform. Many other characters have been added, including significant extensions for the Arabic and Ethiopic scripts, as well as additions to complete the representation of Biblical Hebrew texts. Additional phonetic characters have been included to support phonetic transcriptions other than IPA. Overall character and code ranges are detailed in *Chapter 2, General Structure*.

Note, however, that the Unicode Standard does not encode idiosyncratic, personal, novel, or private-use characters, nor does it encode logos or graphics. Graphologies unrelated to text, such as dance notations, are likewise outside the scope of the Unicode Standard. Font variants are explicitly not encoded. The Unicode Standard reserves 6,400 code points in the BMP for private use, which may be used to assign codes to characters not included in the repertoire of the Unicode Standard. Another 131,068 private-use code points are available outside the BMP, should 6,400 prove insufficient for particular applications.

Standards Coverage

The Unicode Standard is a superset of all characters in widespread use today. It contains the characters from major international and national standards as well as prominent industry character sets. For example, Unicode incorporates the ISO/IEC 6937 and ISO/IEC 8859 families of standards, the SGML standard ISO/IEC 8879, and bibliographic standards such as ISO 5426. Important national standards contained within Unicode include ANSI Z39.64, KS X 1001, JIS X 0208, JIS X 0212, JIS X 0213, GB 2312, GB 18030, HKSCS, and CNS 11643. Industry code pages and character sets from Adobe, Apple, Fujitsu, Hewlett-Packard, IBM, Lotus, Microsoft, NEC, and Xerox are fully represented as well.

For a complete list of ISO and national standards used as sources, see *References*.

The Unicode Standard is fully conformant with the International Standard ISO/IEC 10646:2003, *Information Technology—Universal Multiple-Octet Coded Character Set (UCS)—Architecture and Basic Multilingual Plane, Supplementary Planes,* known as the Universal Character Set (UCS). For more information, see *Appendix C, Relationship to ISO/ IEC 10646*.

New Characters

The Unicode Standard continues to respond to new and changing industry demands by encoding important new characters. As the universal character encoding scheme, the Unicode Standard also responds to scholarly needs. To preserve world cultural heritage, important archaic scripts are encoded as consensus about the encoding is developed.

1.2 Design Goals

The primary goal of the development effort for the Unicode Standard was to remedy two serious problems common to most multilingual computer programs. The first problem was the overloading of the font mechanism when encoding characters. Fonts have often been indiscriminately mapped to the same set of bytes. For example, the bytes 0x00 to 0xFF are often used for both characters and dingbats. The second major problem was the use of multiple, inconsistent character codes because of conflicting national and industry character standards. In Western European software environments, for example, one often finds confusion between the Windows Latin 1 code page 1252 and ISO/IEC 8859-1.

When the Unicode project began in 1988, the groups most affected by the lack of a consistent international character standard included publishers of scientific and mathematical software, newspaper and book publishers, bibliographic information services, and academic researchers. More recently, the computer industry has adopted an increasingly global outlook, building international software that can be easily adapted to meet the needs of particular locations and cultures. The explosive growth of the Internet has added to the demand for a character set standard that can be used all over the world.
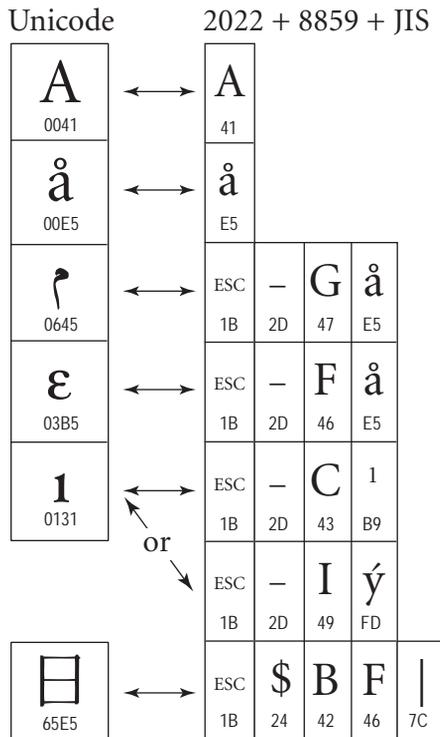
The designers of the Unicode Standard envisioned a uniform method of character identification that would be more efficient and flexible than previous encoding systems. The new system would satisfy the needs of technical and multilingual computing and would encode a broad range of characters for professional-quality typesetting and desktop publishing worldwide.

The Unicode Standard was designed to be:

- *Universal.* The repertoire must be large enough to encompass all characters that are likely to be used in general text interchange, including those in major international, national, and industry character sets.

- *Efficient.* Plain text is simple to parse: software does not have to maintain state or look for special escape sequences, and character synchronization from any point in a character stream is quick and unambiguous. A fixed character code allows for efficient sorting, searching, display, and editing of text.

- *Unambiguous.* Any given Unicode code point always represents the same character.

Figure 1-2 demonstrates some of these features, contrasting the Unicode encoding with mixtures of single-byte character sets with escape sequences to shift the meanings of bytes in the ISO/IEC 2022 framework using multiple character encoding standards.

Figure 1-2. Unicode Compared to the 2022 Framework

Unicode		2022 + 8859 + JIS				
A 0041	⟷	A 41				
å 00E5	⟷	å E5				
؍ 0645	⟷	ESC 1B	– 2D	G 47	å E5	
ε 03B5	⟷	ESC 1B	– 2D	F 46	å E5	
ı 0131	⟷	ESC 1B	– 2D	C 43	1 B9	
	or	ESC 1B	– 2D	I 49	ý FD	
🄳 65E5	⟷	ESC 1B	$ 24	B 42	F 46	| 7C

1.3 Text Handling

Computer text handling involves both encoding and processing. When a word processor user types in the letter "T" via a keyboard, the computer's system software receives a message that the user pressed a key combination for "T", which it encodes as U+0054. The word processor stores the number in memory and also passes it on to the display software responsible for putting the character on the screen. This display software, which may be a windows manager or part of the word processor itself, then uses the number as an index to find an image of a "T", which it draws on the monitor screen. The process continues as the user types in more characters.

The Unicode Standard directly addresses only the encoding and semantics of text and not any other actions performed on the text. In the preceding scenario, the word processor might check the typist's input after it has been encoded to look for misspelled words and then highlight any errors it finds. Alternatively, the word processor might insert line breaks when it counts a certain number of characters entered since the last line break. An important principle of the Unicode Standard is that the standard does not specify how to carry

out these processes as long as the character encoding and decoding is performed properly and the character semantics are maintained.

The difference between identifying a character and rendering it on screen or paper is crucial to understanding the Unicode Standard's role in text processing. The character identified by a Unicode code point is an abstract entity, such as "LATIN CAPITAL LETTER A" or "BENGALI DIGIT FIVE". The mark made on screen or paper, called a glyph, is a visual representation of the character.

The Unicode Standard does not define glyph images. That is, the standard defines how characters are interpreted, not how glyphs are rendered. Ultimately, the software or hardware rendering engine of a computer is responsible for the appearance of the characters on the screen. The Unicode Standard does not specify the precise shape, size, or orientation of on-screen characters.

Text Elements

The successful encoding, processing, and interpretation of text requires appropriate definition of useful elements of text and the basic rules for interpreting text. The definition of text elements often changes depending on the process that handles the text. For example, when searching for a particular word or character written with the Latin script, one often wishes to ignore differences of case. However, correct spelling within a document requires case sensitivity.

The Unicode Standard does not define what is and is not a text element in different processes; instead, it defines elements called *encoded characters*. An encoded character is represented by a number from 0 to $10FFFF_{16}$, called a code point. A text element, in turn, is represented by a sequence of one or more encoded characters.

The Unicode Standard 5.0 – Electronic edition