

# The Unicode Standard

## Version 6.1 – Core Specification

To learn about the latest version of the Unicode Standard, see <http://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.

Copyright © 1991–2012 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 <http://www.unicode.org/reporting.html>. For information about the Unicode terms of use, please see <http://www.unicode.org/copyright.html>.

The Unicode Standard / the Unicode Consortium ; edited by Julie D. Allen ... [et al.]. — Version 6.1.

Includes bibliographical references and index.

ISBN 978-1-936213-02-3 (<http://www.unicode.org/versions/Unicode6.1.0/>)

1. Unicode (Computer character set) I. Allen, Julie D. II. Unicode Consortium.

QA268.U545 2012

ISBN 978-1-936213-02-3

Published in Mountain View, CA

April 2012

# 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 the underpinning for the World Wide Web and the global business environments of today. 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 6.1, 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.

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.

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
l	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

## 1.1 Coverage

The Unicode Standard, Version 6.1, contains 110,116 characters from the world's scripts. These characters are more than sufficient not only for modern communication for the world's languages, but also to represent the classical forms of many languages. The standard includes the European alphabetic scripts, Middle Eastern right-to-left scripts, and scripts of Asia and Africa. Many archaic and historic scripts are encoded. The Han script includes 74,616 ideographic characters defined by national, international, and industry standards of China, Japan, Korea, Taiwan, Vietnam, and Singapore. In addition, the Unicode Standard contains many important symbol sets, including currency symbols, punctuation marks, mathematical symbols, technical symbols, geometric shapes, dingbats, and *emoji*. For overall character and code range information, see *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 10646:2011, *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, 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 Unicode Standard began with a simple goal: to unify the many hundreds of conflicting ways to encode characters, replacing them with a single, universal standard. The pre-existing legacy character encodings were both inconsistent and incomplete—two encodings could use the same codes for two different characters and use different codes for the same characters, while none of the encodings handled any more than a small fraction of the world's languages. Whenever textual data was converted between different programs or platforms, there was a substantial risk of corruption. Programs often were written only to support particular encodings, making development of international versions expensive. As a result, developing countries were particularly hard-hit, as it was not economically feasible to adapt specific versions of programs for smaller markets. Technical fields such as mathematics were also disadvantaged, because they were forced to use special fonts to represent arbitrary characters, often leading to garbled content.

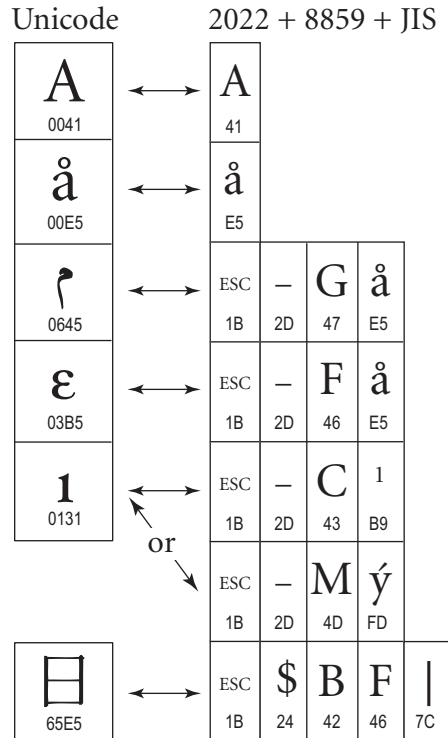
The designers of the Unicode Standard envisioned a uniform method of character identification that would be more efficient and flexible than previous encoding systems. The new system would satisfy the needs of technical and multilingual computing and would encode a broad range of characters for all purposes, including worldwide publication.

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



### 1.3 Text Handling

The assignment of characters is only a small fraction of what the Unicode Standard and its associated specifications provide. The specifications give programmers extensive descriptions and a vast amount of data about the handling of text, including how to:

- divide words and break lines
- sort text in different languages
- format numbers, dates, times, and other elements appropriate to different locales
- display text for languages whose written form flows from right to left, such as Arabic or Hebrew
- display text in which the written form splits, combines, and reorders, such as for the languages of South Asia
- deal with security concerns regarding the many look-alike characters from writing systems around the world

Without the properties, algorithms, and other specifications in the Unicode Standard and its associated specifications, interoperability between different implementations would be

impossible. With the Unicode Standard as the foundation of text representation, all of the text on the Web can be stored, searched, and matched with the same program code.

### ***Characters and Glyphs***

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  $10\text{FFFF}_{16}$ , called a code point. A text element, in turn, is represented by a sequence of one or more encoded characters.

