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

http://www.unicode.org/versions/Unicode5.0.0/bookmarks.html

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

Library of Congress Cataloging-in-Publication Data

The Unicode Standard / the Unicode Consortium ; edited by Julie D. Allen ... [et al.]. — Version 5.0.
p. cm.
Includes bibliographical references and index.
ISBN 0-321-48091-0 (hardcover : alk. paper)
1. Unicode (Computer character set) I. Allen, Julie D.
II. Unicode Consortium.
QA268.U545 2007
005.7'22—dc22

2006023526

ISBN 0-321-48091-0
Text printed in the United States on recycled paper at Courier in Westford, Massachusetts.
First printing, October 2006

Appendix C

Relationship to ISO/IEC 10646

The Unicode Consortium maintains a strong working relationship with ISO/IEC JTC1/SC2/WG2, the working group developing International Standard 10646. Today both organizations are firmly committed to maintaining the synchronization between the Unicode Standard and 10646. Each standard nevertheless uses its own form of reference and, to some degree, separate terminology. This appendix gives a brief history and explains how the standards are related.

C.1 History

Having recognized the benefits of developing a single universal character code standard, members of the Unicode Consortium worked with representatives from the International Organization for Standardization (ISO) during the summer and fall of 1991 to pursue this goal. Meetings between the two bodies resulted in mutually acceptable changes to both Unicode Version 1.0 and the first ISO/IEC Draft International Standard DIS 10646.1, which merged their combined repertoire into a single numerical character encoding. This work culminated in *The Unicode Standard, Version 1.1*.

ISO/IEC 10646-1:1993, *Information Technology—Universal Multiple-Octet*[1] *Coded Character Set (UCS)—Part 1: Architecture and Basic Multilingual Plane*, was published in May 1993 after final editorial changes were made to accommodate the comments of voting members. *The Unicode Standard, Version 1.1*, reflected the additional characters introduced from the DIS 10646.1 repertoire and incorporated minor editorial changes.

Merging *The Unicode Standard, Version 1.0*, and DIS 10646.1 consisted of aligning the numerical values of identical characters and then filling in some groups of characters that were present in DIS 10646.1, but not in the Unicode Standard. As a result, the encoded characters (code points and names) of ISO/IEC 10646-1:1993 and *The Unicode Standard, Version 1.1*, are precisely the same.

1. *Octet* is ISO/IEC terminology for *byte*—that is, an ordered sequence of 8 bits considered as a unit.

Versions 2.0, 2.1, and 3.0 of the Unicode Standard successively added more characters, matching a series of amendments to ISO/IEC 10646-1. *The Unicode Standard, Version 3.0*, is precisely aligned with the second edition of ISO/IEC 10646-1, known as ISO/IEC 10646-1:2000.

In 2001, Part 2 of ISO/IEC 10646 was published as ISO/IEC 10646-2:2001. Version 3.1 of the Unicode Standard was synchronized with that publication, which added supplementary characters for the first time. Subsequently, Versions 3.2 and 4.0 of the Unicode Standard added characters matching further amendments to both parts of ISO/IEC 10646. *The Unicode Standard, Version 4.0*, is precisely aligned with the third version of ISO/IEC 10646 (first edition), published as a single standard merging the former two parts: ISO/IEC 10646:2003.

Versions 4.1 and 5.0 of the Unicode Standard added characters matching Amendments 1 and 2 to ISO/IEC 10646:2003. Version 5.0 also adds four characters for Sindhi support from Amendment 3 to ISO/IEC 10646:2003.

Table C-1 gives the timeline for these efforts.

Table C-1. Timeline

Year	Version	Summary
1989	DP 10646	Draft proposal, independent of Unicode
1990	Unicode Prepublication	Prepublication review draft
1990	DIS-1 10646	First draft, independent of Unicode
1991	Unicode 1.0	Edition published by Addison-Wesley
1992	Unicode 1.0.1	Modified for merger compatibility
1992	DIS-2 10646	Second draft, merged with Unicode
1993	IS 10646-1:1993	Merged standard
1993	Unicode 1.1	Revised to match IS 10646-1:1993
1995	10646 amendments	Korean realigned, plus additions
1996	Unicode 2.0	Synchronized with 10646 amendments
1998	Unicode 2.1	Added euro sign and corrigenda
1999	10646 amendments	Additions
2000	Unicode 3.0	Synchronized with 10646 second edition
2000	IS 10646-1:2000	10646 part 1, second edition, publication with amendments to date
2001	IS 10646-2:2001	10646 part 2 (supplementary planes)
2001	Unicode 3.1	Synchronized with 10646 part 2
2002	Unicode 3.2	Synchronized with Amd 1 to 10646 part 1
2003	Unicode 4.0	Synchronized with 10646 third version
2003	IS 10646:2003	10646 third version (first edition), merging the two parts
2005	Unicode 4.1	Synchronized with Amd 1 to 10646:2003
2006	Unicode 5.0	Synchronized with Amd 2 to 10646:2003, plus Sindhi additions

Unicode 1.0

The combined repertoire presented in ISO/IEC 10646 is a superset of *The Unicode Standard, Version 1.0,* repertoire as amended by *The Unicode Standard, Version 1.0.1*. *The Unicode Standard, Version 1.0*, was amended by the *Unicode 1.0.1 Addendum* to make the Unicode Standard a proper subset of ISO/IEC 10646. This effort entailed both moving and eliminating a small number of characters.

Unicode 2.0

The Unicode Standard, Version 2.0, covered the repertoire of *The Unicode Standard, Version 1.1* (and IS 10646), plus the first seven amendments to IS 10646, as follows:

Amd. 1: UTF-16

Amd. 2: UTF-8

Amd. 3: Coding of C1 Controls

Amd. 4: Removal of Annex G: UTF-1

Amd. 5: Korean Hangul Character Collection

Amd. 6: Tibetan Character Collection

Amd. 7: 33 Additional Characters (Hebrew, Long S, Dong)

In addition, *The Unicode Standard, Version 2.0*, covered Technical Corrigendum No. 1 (on renaming of AE LIGATURE to LETTER) and such Editorial Corrigenda to ISO/IEC 10646 as were applicable to the Unicode Standard. The euro sign and the object replacement character were added in Version 2.1, per amendment 18 of ISO 10646-1.

Unicode 3.0

The Unicode Standard, Version 3.0, is synchronized with the second edition of ISO/IEC 10646-1. The latter contains all of the published amendments to 10646-1; the list includes the first seven amendments, plus the following:

Amd. 8: Addition of Annex T: Procedure for the Unification and Arrangement of CJK Ideographs

Amd. 9: Identifiers for Characters

Amd. 10: Ethiopic Character Collection

Amd. 11: Unified Canadian Aboriginal Syllabics Character Collection

Amd. 12: Cherokee Character Collection

Amd. 13: CJK Unified Ideographs with Supplementary Sources (Horizontal Extension)

Amd. 14: Yi Syllables and Yi Radicals Character Collection

Amd. 15: Kangxi Radicals, Hangzhou Numerals Character Collection

Amd. 16: Braille Patterns Character Collection

Amd. 17: CJK Unified Ideographs Extension A (Vertical Extension)

Amd. 18: Miscellaneous Letters and Symbols Character Collection (which includes the euro sign)

Amd. 19: Runic Character Collection

Amd. 20: Ogham Character Collection

Amd. 21: Sinhala Character Collection

Amd. 22: Keyboard Symbols Character Collection

Amd. 23: Bopomofo Extensions and Other Character Collection

Amd. 24: Thaana Character Collection

Amd. 25: Khmer Character Collection

Amd. 26: Myanmar Character Collection

Amd. 27: Syriac Character Collection

Amd. 28: Ideographic Description Characters

Amd. 29: Mongolian

Amd. 30: Additional Latin and Other Characters

Amd. 31: Tibetan Extension

The second edition of 10646-1 also contains the contents of Technical Corrigendum No. 2 and all the Editorial Corrigenda to the year 2000.

Unicode 4.0

The Unicode Standard, Version 4.0, is synchronized with the third version of ISO/IEC 10646. The third version of ISO/IEC 10646 is the result of the merger of the second edition of Part 1 (ISO/IEC 10646-1:2000) with the first edition of Part 2 (ISO/IEC 10646-2:2001) into a single publication. The third version incorporates the published amendments to 10646-1 and 10646-2:

Amd. 1 (to part 1): Mathematical symbols and other characters

Amd. 2 (to part 1): Limbu, Tai Le, Yijing, and other characters

Amd. 1 (to part 2): Aegean, Ugaritic, and other characters

The third version of 10646 also contains all the Editorial Corrigenda to date.

Unicode 5.0

The Unicode Standard, Version 5.0, is synchronized with ISO/IEC 10646:2003 plus its first two published amendments:

Amd. 1: Glagolitic, Coptic, Georgian and other characters

Amd. 2: N'Ko, Phags-Pa, Phoenician and Cuneiform

Four Devanagari characters for the support of the Sindhi language (U+097B, U+097C, U+097E, U+097F) were added in Version 5.0 per Amendment 3 of ISO 10646.

The synchronization of *The Unicode Standard, Version 5.0*, with the third version of ISO/IEC 10646 plus its amendments means that the repertoire, encoding, and names of all characters are identical between the two standards at those version levels, and that all other material from the amendments to 10646 that have a bearing on the text of the Unicode Standard have been taken into account in the revision of the Unicode Standard.

C.2 Encoding Forms in ISO/IEC 10646

ISO/IEC 10646 defines four alternative forms of encoding: UCS-4, UCS-2, UTF-8, and UTF-16. UTF-8 and UTF-16 are discussed in *Section C.3, UCS Transformation Formats*.

UCS-4. UCS-4 stands for "Universal Character Set coded in 4 octets" and is considered the canonical form for 10646. It is a four-octet (32-bit) encoding containing 2^{31} code positions. These code positions are conceptually divided into 128 *groups* of 256 *planes*, with each plane containing 256 *rows* of 256 *cells*.

ISO/IEC 10646 states that all future assignments of characters to 10646 will be allocated on the BMP or the first 14 supplementary planes. This is to ensure interoperability between the UCS transformation formats (see below). It also guarantees interoperability with implementations of the Unicode Standard, for which only code positions $0..10FFFF_{16}$ are meaningful. The former provision for private-use code positions in groups 60 to 7F and in planes E0 to FF in 10646 has been removed from 10646. As a consequence, UCS-4 can now be taken effectively as an alias for the Unicode encoding form UTF-32, except that UTF-32 has the extra requirement that additional Unicode semantics be observed for all characters.

UCS-2. UCS-2 stands for "Universal Character Set coded in 2 octets" and is also known as "the two-octet BMP form." It is the two-octet (16-bit) encoding consisting only of code positions for plane zero, the *Basic Multilingual Plane*.

Zero Extending

The character "A", U+0041 LATIN CAPITAL LETTER A, has the unchanging numerical value 41 hexadecimal. This value may be extended by any quantity of leading zeros to serve in the context of the following encoding standards and transformation formats (see *Table C-2*).

This design eliminates the problem of disparate values in all systems that use either of the standards and their transformation formats.

Table C-2. Zero Extending

Bits	*Standard*	*Binary*	*Hex*	*Dec*	*Char*
7	ASCII	1000001	41	65	A
8	8859-1	01000001	41	65	A
16	UTF-16, UCS-2	00000000 01000001	41	65	A
32	UTF-32, UCS-4	00000000 00000000 00000000 01000001	41	65	A

C.3 UCS Transformation Formats

UTF-8

The term *UTF-8* in ISO/IEC 10646 stands for "UCS Transformation Format, 8-bit form." UTF-8 is an alternative coded representation form for all of the characters of ISO/IEC 10646. The ISO/IEC definition is identical in format to UTF-8 as described under definition D92 in *Section 3.9, Unicode Encoding Forms*.

UTF-8 can be used to transmit text data through communications systems that assume that individual octets in the range of x00 to x7F have a definition according to ISO/IEC 4873, including a C0 set of control functions according to the 8-bit structure of ISO/IEC 2022. UTF-8 also avoids the use of octet values in this range that have special significance during the parsing of file name character strings in widely used file-handling systems.

The definition of UTF-8 in Annex D of ISO/IEC 10646:2003 also allows for the use of five- and six-byte sequences to encode characters that are outside the range of the Unicode character set; those five- and six-byte sequences are illegal for the use of UTF-8 as an encoding form of Unicode characters. ISO/IEC 10646 does not allow mapping of surrogate code positions, known as RC-elements in that standard; that restriction is identical to the restriction for the Unicode definition of UTF-8.

UTF-16

The term *UTF-16* in ISO/IEC 10646 stands for "UCS Transformation Format for 16 Planes of Group 00." It is defined in Annex C of ISO/IEC 10646:2003. In UTF-16, each BMP code position represents itself. Non-BMP code positions of ISO/IEC 10646 in planes 1 to 16 are represented using pairs of special codes. UTF-16 defines the transformation between the UCS-4 code positions in planes 1 to 16 of Group 00 and the pairs of special codes and is identical to the UTF-16 encoding form defined in the Unicode Standard under definition D91 in *Section 3.9, Unicode Encoding Forms*.

In ISO/IEC 10646, *high-surrogates* are called "RC-elements from the high-half zone" and *low-surrogates* are called "RC-elements from the low-half zone." Together, they constitute the S (Special) Zone of the BMP.

UTF-16 represents the BMP and the next 16 planes.

C.4 Synchronization of the Standards

Programmers and system users should treat the encoded character values from the Unicode Standard and ISO/IEC 10646 as identities, especially in the transmission of raw character data across system boundaries. The Unicode Consortium and ISO/IEC JTC1/SC2/WG2 are committed to maintaining the synchronization between the two standards.

However, the Unicode Standard and ISO/IEC 10646 differ in the precise terms of their conformance specifications. Any Unicode implementation will conform to ISO/IEC 10646, level 3, but because the Unicode Standard imposes additional constraints on character semantics and transmittability, not all implementations that are compliant with ISO/IEC 10646 will be compliant with the Unicode Standard.

C.5 Identification of Features for the Unicode Standard

ISO/IEC 10646 provides mechanisms for specifying a number of implementation parameters, generating what may be termed instantiations of the standard. ISO/IEC 10646 contains no means of explicitly declaring the Unicode Standard as such. As a whole, however, the Unicode Standard may be considered as encompassing the entire repertoire of ISO/IEC 10646 and having the following features (as well as additional semantics):

- Numbered subset 307 (UNICODE 5.0)
- UTF-8, UTF-16, or UCS-4 (= UTF-32)
- Implementation level 3 (allowing both combining marks and precomposed characters)
- Device type 1 (receiving device with full retransmission capability)

Few applications are expected to make use of all of the characters defined in ISO/IEC 10646. The conformance clauses of the two standards address this situation in very different ways. ISO/IEC 10646 provides a mechanism for specifying included subsets of the character repertoire, permitting implementations to ignore characters that are not included (see normative Annex A of ISO/IEC 10646). A Unicode implementation requires a minimal level of handling all character codes—namely, the ability to store and retransmit them undamaged. Thus the Unicode Standard encompasses the entire ISO/IEC 10646 repertoire without requiring that any particular subset be implemented.

The Unicode Standard does not provide formal mechanisms for identifying a stream of bytes as Unicode characters, although to some extent this function is served by use of the *byte order mark* (U+FEFF) to indicate byte ordering. ISO/IEC 10646 defines an ISO/IEC 2022 control sequence to introduce the use of 10646. ISO/IEC 10646 also allows the use of U+FEFF as a "signature" as described in ISO/IEC 10646. This optional "signature" convention for identification of UTF-8, UTF-16, and UCS-4 is described in the informative Annex

H of 10646. It is consistent with the description of the *byte order mark* in *Section 16.8, Specials*.

C.6 Character Names

Unicode character names follow the ISO/IEC character naming guidelines (summarized in informative Annex L of ISO/IEC 10646). In the first version of the Unicode Standard, the naming convention followed the ISO/IEC naming convention, but with some differences that were largely editorial. For example,

ISO/IEC 10646 name	029A	LATIN SMALL LETTER CLOSED OPEN E
Unicode 1.0 name	029A	LATIN SMALL LETTER CLOSED EPSILON

In the ISO/IEC framework, the unique character name is viewed as the major resource for both character semantics and cross-mapping among standards. In the framework of the Unicode Standard, character semantics are indicated via character properties, functional specifications, usage annotations, and name aliases; cross-mappings among standards are provided in the form of explicit tables available on the Unicode Web site. The disparities between the Unicode 1.0 names and ISO/IEC 10646 names have been remedied by adoption of ISO/IEC 10646 names in the Unicode Standard. The names adopted by the Unicode Standard are from the English-language version of ISO/IEC 10646, even when other language versions are published by ISO.

C.7 Character Functional Specifications

The core of a character code standard is a mapping of code points to characters, but in some cases the semantics or even the identity of the character may be unclear. Certainly a character is not simply the representative glyph used to depict it in the standard. For this reason, the Unicode Standard supplies the information necessary to specify the semantics of the characters it encodes.

Thus the Unicode Standard encompasses far more than a chart of code points. It also contains a set of extensive character functional specifications and data, as well as substantial background material designed to help implementers better understand how the characters interact. The Unicode Standard specifies properties and algorithms. Conformant implementations of the Unicode Standard will also be conformant with ISO/IEC 10646, level 3.

Compliant implementations of ISO/IEC 10646 can be conformant to the Unicode Standard—as long as the implementations conform to all additional specifications that apply to the characters of their adopted subsets, and as long as they support all Unicode characters outside their adopted subsets in the manner referred to in *Section C.5, Identification of Features for the Unicode Standard*.

