The Unicode Standard
Version 8.0 – Core Specification

To learn about the latest version of the Unicode Standard, see http://www.unicode.org/versions/latest/.

The Unicode Standard / the Unicode Consortium ; edited by Julie D. Allen ... [et al.]. — Version 8.0
Includes bibliographical references and index.
ISBN 978-1-936213-10-8 (http://www.unicode.org/versions/Unicode8.0.0/)
1. Unicode (Computer character set) I. Allen, Julie D. II. Unicode Consortium.
QA268.U545 2015

ISBN 978-1-936213-10-8
Published in Mountain View, CA
August 2015

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 ISO/IEC 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 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.

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 added four characters for Sindhi support from Amendment 3 to ISO/IEC 10646:2003. Version 5.1 added the rest of the characters from Amendment 3 and all of the characters from Amendment 4 to ISO/IEC 10646:2003. Version 5.2 added all of the characters from Amendments 5 and 6 to ISO/IEC 10646:2003. Version 6.0 added all of the characters from Amendments 7 and 8 to ISO/IEC 10646:2003.

In 2010, ISO approved republication of ISO/IEC 10646 as a second edition, ISO/IEC 10646:2011, consolidating all of the contents of Amendments 1 through 8 to the 2003 first edition. *The Unicode Standard, Version 6.0* is aligned with that second edition of the International Standard, with the addition of U+20B9 INDIAN RUPEE SIGN, accelerated into Version 6.0 based on approval for the third edition of ISO/IEC 10646.

The Unicode Standard, Version 6.1 is aligned with the third edition of the International Standard: ISO/IEC 10646:2012. The third edition was approved for publication without an

intervening amendment to the second edition. *The Unicode Standard, Version 6.2* added a single character, U+20BA TURKISH LIRA SIGN. Version 6.3 added five more characters, including new bidirectional format controls.

The Unicode Standard, Version 7.0 is aligned with Amendments 1 and 2 to ISO/IEC 10646:2012. Those amendments include the six characters which were added in Version 6.2 and Version 6.3, as well as many others. Version 7.0 also includes U+20BD RUBLE SIGN, accelerated into Version 7.0 based on approval for the fourth edition of ISO/IEC 10646.

The Unicode Standard, Version 8.0 is aligned with Amendment 1 of ISO/IEC 10646:2014, the fourth edition of ISO/IEC 10646. Version 8.0 also includes U+20BE LARI SIGN, nine additional CJK unified ideographs, and 41 *emoji* characters, based on approval for Amendment 2 to the fourth edition of ISO/IEC 10646.

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
2008	Unicode 5.1	Synchronized with Amd 3 and Amd 4 to 10646:2003
2009	Unicode 5.2	Synchronized with Amd 5 and Amd 6 to 10646:2003

Table C-1. Timeline (Continued)

Year	Version	Summary
2010	Unicode 6.0	Synchronized with 10646 second edition of third version, plus the Indian rupee sign
2011	IS 10646:2011	10646 second edition of third version
2012	Unicode 6.1	Synchronized with 10646 third edition of third version
2012	IS 10646:2012	10646 third edition of third version
2012	Unicode 6.2	Added Turkish lira sign
2013	Unicode 6.3	Added several bidirectional controls
2014	Unicode 7.0	Synchronized with Amd 1 and Amd 2 to 10646:2012, plus the ruble sign
2015	Unicode 8.0	Synchronized with Amd 1 to 10646:2014, plus 51 additional characters

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/IEC 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 ISO/IEC 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 ISO/IEC 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/IEC 10646.

Unicode 6.0

The Unicode Standard, Version 6.0, is synchronized with the second edition of ISO/IEC 10646. The second edition of the third version of ISO/IEC 10646 consolidates all of the repertoire additions from the published eight amendments of ISO/IEC 10646:2003. These include the first two amendments listed under Unicode 5.0, plus the following:

Amd. 3: Lepcha, Ol Chiki, Saurashtra, Vai, and other characters

Amd. 4: Cham, Game Tiles, and other characters

Amd. 5: Tai Tham, Tai Viet, Avestan, Egyptian Hieroglyphs, CJK Unified Ideographs Extension C, and other characters

Amd. 6: Javanese, Lisu, Meetei Mayek, Samaritan, and other characters

Amd. 7: Mandaic, Batak, Brahmi, and other characters

Amd. 8: Additional symbols, Bamum supplement, CJK Unified Ideographs Extension D, and other characters

One additional character, for the support of the new Indian currency symbol (U+20B9 INDIAN RUPEE SIGN), was accelerated into Version 6.0, based on its approval for the third edition of ISO/IEC 10646.

Unicode 7.0

The Unicode Standard, Version 7.0, is synchronized with the third edition of ISO/IEC 10646 plus its two published amendments:

Amd. 1: Linear A, Palmyrene, Manichaean, Khojki, Khudawadi, Bassa Vah, Duployan, and other characters

Amd. 2: Caucasian Albanian, Psalter Pahlavi, Mahajani, Grantha, Modi, Pahawh Hmong, Mende Kikakui, and other characters

One additional character, for the support of the new Russian currency symbol (U+20BD RUBLE SIGN), was accelerated into Version 7.0, based on its approval for the fourth edition of ISO/IEC 10646.

Unicode 8.0

The Unicode Standard, Version 8.0, is synchronized with the fourth edition of ISO/IEC 10646, plus its first published amendment:

Amd. 1: Cherokee supplement and other characters

An additional 51 characters were accelerated into Version 8.0, based on their approval for Amendment 2 to the fourth edition of ISO/IEC 10646. These include U+20BE LARI SIGN, for the support of the new Georgian currency symbol, nine additional CJK unified ideographs, and 41 *emoji* characters.

The synchronization of *The Unicode Standard, Version 8.0*, with ISO/IEC 10646:2014 plus its published amendment means that the repertoire, encoding, and names of all characters are identical between the two standards at those version levels, except for the 51 additional characters from Amendment 2 which were accelerated for publication in the Unicode Standard. All other changes to the text of 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 10646:2011 has significantly revised its discussion of encoding forms, compared to earlier editions of that standard. The terminology for encoding forms (and encoding schemes) in 10646 now matches exactly the terminology used in the Unicode Standard. Furthermore, 10646 is now described in terms of a codespace U+0000..U+10FFFF, instead of a 31-bit codespace, as in earlier editions. This convergence in codespace description has eliminated any discrepancies in possible interpretation of the numeric values greater than 0x10FFFF. As a result, this section now merely notes a few items of mostly historic interest regarding encoding forms and terminology.

UCS-4. UCS-4 stands for "Universal Character Set coded in 4 octets." It is now treated simply as a synonym for UTF-32, and is considered the canonical form for representation of characters in 10646.

UCS-2. UCS-2 stands for "Universal Character Set coded in 2 octets" and is also known as "the two-octet BMP form." It was documented in earlier editions of 10646 as the two-octet (16-bit) encoding consisting only of code positions for plane zero, the *Basic Multilingual Plane*. This documentation has been removed from ISO/IEC 10646:2011, and the term UCS-2 should now be considered obsolete. It no longer refers to an encoding form in either 10646 or the Unicode Standard.

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*).

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	00000000 01000001	41	65	A
32	UTF-32, UCS-4	00000000 00000000 00000000 01000001	41	65	A

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

C.3 UTF-8 and UTF-16

UTF-8

The ISO/IEC 10646 definition of UTF-8 is identical 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.

UTF-16

The ISO/IEC 10646 definition of UTF-16 is identical to UTF-16 as described under Definition D91 in *Section 3.9, Unicode Encoding Forms*.

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, 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 Unicode

ISO/IEC 10646 provides mechanisms for specifying a number of implementation parameters. 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 collection 315 (UNICODE 8.0)
- Encoding forms: UTF-8, UTF-16, or UTF-32
- Encoding schemes: UTF-8, UTF-16BE, UTF-16LE, UTF-16, UTF-32BE, UTF-32LE, or UTF-32

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 UTF-32 is described in the informative Annex H of 10646. It is consistent with the description of the *byte order mark* in *Section 23.8, 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 website. 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.

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 Unicode*.