

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-Octet¹ 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 character code values of ISO/IEC 10646-1:1993 UCS-2 and *The Unicode Standard, Version 1.1* are precisely the same. Versions 2.0 and 3.0 of the Unicode Standard have added more characters, matching recent additions to ISO/IEC 10646-1:2000. *Table C-1* gives the timeline for these efforts.

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

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 (expected)	IS 10646-1:2000	10646 part 1, second edition, publication with amendments to date

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* also 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 date.

The synchronization of *The Unicode Standard, Version 3.0* with the second edition of ISO/IEC 10646-1 means that the repertoire, encoding, and names of all characters are identical between the two standards, and that all other material from the amendments to 10646-1 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 two alternative forms of encoding:

- A four-octet (32-bit) encoding containing 2^{31} code positions. These code positions are conceptually divided into 128 *groups* of 256 *planes*, each plane containing 256 *rows* of 256 *cells*.
- A two-octet (16-bit) encoding consisting of plane zero, the *Basic Multilingual Plane* (BMP).

The 32-bit form is referred to as UCS-4 (Universal Character Set coded in 4 octets) and the 16-bit form is referred to as UCS-2 (Universal Character Set coded in 2 octets).

The code numbers from 0 through 65,535 decimal (0–FFFF hexadecimal) can be represented by character code values of 16 bits. The most useful characters (that is, the characters found in major existing standards worldwide) are assigned in the BMP. ISO/IEC 10646

does not currently define any characters in other planes. The encoding of characters in other planes will be covered by ISO/IEC 10646-2.

ISO/IEC 10646-1 does not currently encode any characters outside of the BMP. The character repertoires and encoding assignments of the Unicode Standard and ISO/IEC 10646-1 are identical.

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 fixed-length encoding standards (see *Table C-2*).

Table C-2. Zero Extending

<i>Bits</i>	<i>Standard</i>	<i>Binary</i>	<i>Hex</i>	<i>Dec</i>	<i>Char</i>
7	ASCII	1000001	41	65	A
8	8859-1	01000001	41	65	A
16	Unicode, UCS-2	00000000 01000001	41	65	A
32	UCS-4	00000000 00000000 00000000 01000001	41	65	A

This design eliminates the problem of disparate code values in all systems that use any of the standards just mentioned.

C.3 UCS Transformation Formats

UTF-8

The term *UTF-8* 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 in *Section 2.3, 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-8 was first defined in ISO/IEC 10646 Amendment 2 and will be included in the next edition of ISO/IEC 10646-1.

UTF-16

The term *UTF-16* stands for UCS Transformation Format for 16 Planes of Group 00. UTF-16 is the ISO/IEC encoding that is equivalent to the Unicode Standard with the use of surrogates as described in *Chapter 3, Conformance*. In UTF-16, each UCS-2 code value represents itself. Non-BMP code values of ISO/IEC 10646 in planes 1..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 transformation defined in the Unicode Standard under D.28 in *Section 3.7, Surrogates*. Sample code for transforming UCS-4 into the Unicode Standard with surrogates is located on the CD-ROM.

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. This system should not be an undue limitation because ISO JTC1/SC2/WG2 has no intention of assigning characters outside of planes 1..14 as it would break synchronization with the Unicode Standard. Planes 15 and 16 (000F0000..000FFFFF₁₆ and 00100000..0010FFFF₁₆) are reserved for private use. The UCS-4 private-use code values in groups 60 to 7F and in planes E0 to FF in group 00 are not accessible using UTF-16. Use of these private-use code values is strongly discouraged because data encoded with these code values will not be interchangeable with Unicode implementations. Planes 15 and 16 should be used instead.

Applications interchanging ISO/IEC 10646 data containing non-BMP code values in planes 1..16 of ISO/IEC 10646 should use UTF-16 as the default encoding form in the absence of information to the contrary.

C.4 Synchronization of the Standards

The goal of merging the Unicode Standard and DIS 10646.1 has been realized, making character code assignments *identical* in the Unicode Standard and ISO/IEC 10646. Programmers and system users should treat the character code values from the Unicode Standard and ISO/IEC 10646 as identities, especially in the transmission of raw character data across system boundaries.

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 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 300 (BMP)
- UTF-16 (if surrogates are used; UCS-2 otherwise)
- 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 the ISO/IEC 10646 Basic Multilingual Plane. 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 Informative 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 Basic Multilingual Plane 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 discerning between forms UCS-2 and UCS-4 is brought to the attention of Unicode implementers. This method is summarized in *Section 13.6, Specials*.

C.6 Character Names

Unicode character names follow the ISO/IEC character naming guidelines (summarized in Informative Annex K of ISO/IEC 10646). In the prior version of the Unicode Standard, the naming convention followed the ISO/IEC naming convention,¹ 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 alias names, usage annotations, character properties, and functional specifications as mentioned in *Chapter 3, Conformance*; cross-mappings among standards are provided in the form of explicit tables. The disparities between the Unicode Standard, Version 1.0 names and ISO/IEC 10646 names have been remedied by adoption of ISO/IEC 10646 names in the Unicode Standard. If the Unicode Standard, Version 1.0 name differed from the ISO/IEC 10646 name, then the previous name is provided as an alias in the Unicode Character Database.

C.7 Character Functional Specifications

The core of a character code standard is a mapping of code values 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 necessary information to specify the semantics of the characters it encodes.

Thus the Unicode Standard consists of far more than a chart of code values. 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*.

1. 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.

This PDF file is an excerpt from *The Unicode Standard, Version 3.0*, issued by the Unicode Consortium and published by Addison-Wesley. The material has been modified slightly for this online edition, however the PDF files have not been modified to reflect the corrections found on the Updates and Errata page (see <http://www.unicode.org/unicode/uni2errata/UnicodeErrata.html>). More recent versions of the Unicode standard exist (see <http://www.unicode.org/unicode/standard/versions/>).

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 Addison-Wesley was aware of a trademark claim, the designations have been printed in initial capital letters. However, not all words in initial capital letters are trademark designations.

The authors and publisher have taken care in 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.

ISBN 0-201-61633-5

Copyright © 1991-2000 by Unicode, Inc.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior written permission of the publisher or Unicode, Inc.

This book is set in Minion, designed by Rob Slimbach at Adobe Systems, Inc. It was typeset using FrameMaker 5.5 running under Windows NT. ASMUS, Inc. created custom software for chart layout. The Han radical-stroke index was typeset by Apple Computer, Inc. The following companies and organizations supplied fonts:

Apple Computer, Inc.
Atelier Fluxus Virus
Beijing Zhong Yi (Zheng Code) Electronics Company
DecoType, Inc.
IBM Corporation
Monotype Typography, Inc.
Microsoft Corporation
Peking University Founder Group Corporation
Production First Software

Additional fonts were supplied by individuals as listed in the *Acknowledgments*.

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.

All other company and product names are trademarks or registered trademarks of the company or manufacturer, respectively.

The publisher offers discounts on this book when ordered in quantity for special sales. For more information please contact:

Corporate, Government, and Special Sales
Addison Wesley Longman, Inc.
One Jacob Way
Reading, Massachusetts 01867

Visit A-W on the Web: <http://www.awl.com/cseng/>

First printing, January 2000.