INTERNATIONAL STANDARD

ISO/IEC 10646

Fifth edition
2017-12

Information technology — Universal Coded Character Set (UCS)

Technologies de l'information — Jeu universel de caractères codés (JUC)

Reference number
ISO/IEC 10646:2017(E)

ISO copyright office
Ch. de Blandonnet 8 • CP 401
CH-1214 Vernier, Geneva, Switzerland
Tel. +41 22 749 01 11
Fax +41 22 749 09 47
copyright@iso.org
www.iso.org

CONTENTS

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation on the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see the following URL: www.iso.org/iso/foreword.html.

The committee responsible for this document is ISO/IEC JTC 1, *Information technology, SC 2, Coded character sets.*

This fifth edition of ISO/IEC 10646 cancels and replaces the fourth edition (ISO/IEC 10646:2014), which has been technically revised. It also incorporates ISO/IEC 10646:2014/Amd 1:2015 and ISO/IEC 10646:2014/Amd 2:2016.

This edition includes the following significant changes with respect to the previous edition:

- New scripts covered: Adlam, Bhaiksuki, , Marchen, Masaram Gondhi, Newa, Nushu, Osage, Soyombo, Tangut, and Zanabazar Square,
- Existing scripts significantly extended: Cherokee, CJK Unified Ideographs (Extension F),
- New Emoji symbols.

Introduction

This International Standard specifies the Universal Coded Character Set (UCS). It is applicable to the representation, transmission, interchange, processing, storage, input and presentation of the written form of the languages of the world as well as additional symbols.

By defining a consistent way of encoding multilingual text it enables the exchange of data internationally. The information technology industry gains data stability, greater global interoperability and data interchange. This International Standard has been widely adopted in new Internet protocols and implemented in modern operating systems and computer languages. This edition covers over 130 000 characters from the world's scripts.

Information technology — Universal Coded Character Set (UCS)

1 Scope

This International Standard specifies the Universal Coded Character Set (UCS). It is applicable to the representation, transmission, interchange, processing, storage, input, and presentation of the written form of the languages of the world as well as of additional symbols.

This International Standard

- specifies the architecture of this International Standard,
- defines terms used in this International Standard,
- describes the general structure of the UCS codespace,
- specifies the Basic Multilingual Plane (BMP) of the UCS,
- specifies supplementary planes of the UCS: the Supplementary Multilingual Plane (SMP), the Supplementary Ideographic Plane (SIP), the Tertiary Ideographic Plane (TIP), and the Supplementary Special-purpose Plane (SSP),
- defines a set of graphic characters used in scripts and the written form of languages on a world-wide scale,
- specifies the names for the graphic characters and format characters of the BMP, SMP, SIP, TIP, SSP and their coded representations within the UCS codespace,
- specifies the coded representations for control characters and private use characters,
- specifies three encoding forms of the UCS: UTF-8, UTF-16, and UTF-32,
- specifies seven encoding schemes of the UCS: UTF-8, UTF-16, UTF-16BE, UTF-16LE, UTF-32, UTF-32BE, and UTF-32LE,
- specifies the management of future additions to this coded character set.

The UCS is an encoding system different from that specified in ISO/IEC 2022. The method to designate UCS from ISO/IEC 2022 is specified in 12.2.

A graphic character will be assigned only one code point in the standard, located either in the BMP or in one of the supplementary planes.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 2022:1994 *Information technology — Character code structure and extension techniques.*

ISO/IEC 6429:1992 *Information technology — Control functions for coded character sets.*

Unicode Standard Annex, UAX #9, *The Unicode Bidirectional Algorithm:*
http://www.unicode.org/reports/tr9/tr9-35.html

Unicode Standard Annex, UAX #15, *Unicode Normalization Forms:*
http://www.unicode.org/reports/tr15/tr15-44.html

Unicode Technical Standard, *UTS #37, Ideographic Variation Database:*
http://www.unicode.org/reports/tr37/tr37-8.html

Unicode Standard Version 9.0, *Chapter 4, Character Properties*
http://www.unicode.org/versions/Unicode9.0.0/ch04.pdf
Section 4.3, Combining Classes – Normative
Section 4.5, General Category – Normative
Section 4.7, Bidi Mirrored – Normative

Unicode Standard Version 9.0, *Age Property:*
http://www.unicode.org/Public/9.0.0/ucd/DerivedAge.txt

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— IEC Electropedia: available at http://www.electropedia.org/

— ISO Online browsing platform: available at http://www.iso.org/obp

3.1
base character
graphic character which is not a combining character

Note 1 to entry – Most graphic characters are base characters. This sense of graphic combination does not preclude the presentation of base characters from adopting different contextual forms or from participating in ligatures.

Note 2 to entry – A base character typically does not graphically combine with preceding characters. There are exceptions for some complex writing systems.

3.2
Basic Multilingual Plane
BMP
plane 00 of the UCS codespace

3.3
block
contiguous range of code points to which a set of characters that share common characteristics, such as a script, are allocated; a block does not overlap another block; one or more of the code points within a block may have no character allocated to them

3.4
canonical form
form with which characters of this coded character set are specified using a single code point within the UCS codespace

Note 1 to entry – The canonical form is not to be confused with an encoding form which describes the relationship between UCS code points and one or several code units (see 3.23).

3.5
character
member of a set of elements used for the organization, control, or representation of textual data

Note 1 to entry – A graphic symbol can be represented by a sequence of one or several coded characters.

3.6
character boundary
(code unit sequence) demarcation between the last code unit of a coded character and the first code unit of the next coded character

3.7
code chart
code table
rectangular array showing the representation of coded characters allocated within a range of the UCS codespace

3.8
coded character
association between a character and a code point

3.9
coded character set
set of coded characters

3.10
code point
code position
value in the UCS codespace

3.11
code unit
minimal bit combination that can represent a unit of encoded text for processing or interchange

Note 1 to entry – Examples of code units are octets (8-bit code units) used in the UTF-8 encoding form, 16-bit code units in the UTF-16 encoding form, and 32-bit code units in the UTF-32 encoding form.

3.12
code unit sequence
CC-data-element
coded-character-data-element
element of interchanged information that is specified to consist of a sequence of code units, in accordance with one or more identified standards for coded character sets

Note 1 to entry – Such sequence can contain code units associated with any type of code points.

Note 2 to entry – Since its second edition, ISO/IEC 10646:2011, this International Standard does not use implementation levels. Its definition of code unit sequence corresponds to the former unrestricted implementation level 3. Other definitions of code unit sequence, previously known as level 1 and 2, are deprecated. To maintain compatibility with these previous editions, in the context of identification of coded representation in International Standards such as ISO/IEC 8824 and ISO/IEC 8825, the concept of implementation level can still be referenced as 'Implementation level 3'. See Annex N.

3.13
collection
numbered and named set of entities

Note 1 to entry – For a non extended collection, these entities consist only of those coded characters whose code points lie within one or more identified ranges (see also 3.25 for extended collection).

Note 2 to entry – If any of the identified ranges include code points to which no character is allocated, the repertoire of the collection will change if an additional character is assigned to any of those code points at a future amendment of this International Standard. However, it is intended that the collection number and name will remain unchanged in future editions of this International Standard.

3.14
combining character
character which has General Category values of Spacing Combining Mark (Mc), Non Spacing Mark (Mn), and Enclosing Mark (Me)

Note 1 to entry – These characters are intended for combination with the preceding non-combining graphic character, or with a sequence of combining characters preceded by a non-combining character (see also 3.17).

3.15
combining class
value associated with each combining character determining its typographical interaction and its canonical ordering within a sequence of combining characters

3.16
compatibility character
graphic character included as a coded character of this International Standard primarily for compatibility with existing coded character sets

3.17
composite sequence
sequence of graphic characters consisting of a base character followed by one or more combining characters, ZERO WIDTH JOINER, or ZERO WIDTH NON-JOINER (see also 3.14)

Note 1 to entry – A graphic symbol for a composite sequence generally consists of the combination of the graphic symbols of each character in the sequence.

Note 2 to entry – A composite sequence can be used to represent characters not encoded in the repertoire of this International Standard.

3.18
control character
control function the coded representation of which consists of a single code point

Note 1 to entry – Although control characters are often 'named' using terms such as DELETE, FORM FEED, ESC, these qualifiers do not correspond to formal character names. See Clause 11 for a list of the long names used by ISO/IEC 6429 in association with the control characters.

3.19
control function
action that affects the recording, processing, transmission, or interpretation of data, and that is represented by a code unit sequence

3.20
decomposition mapping
mapping from a character to a sequence of one or more characters that is a canonical or compatibility equivalent

3.21
default state
state that is assumed when no state has been explicitly specified (see F.2.1, F.2.2, and F.2.3)

3.22
device
component of information processing equipment which can transmit and/or receive coded information within code unit sequences

Note 1 to entry – It may be an input/output device in the conventional sense, or a process such as an application program or gateway function.

3.23
encoding form
form that determines how each UCS code point for a UCS character is to be expressed as one or more code units used by the encoding form

Note 1 to entry – This International Standard specifies UTF-8, UTF-16, and UTF-32.

3.24
encoding scheme
scheme that specifies the serialization of the code units from the encoding form into octets

Note 1 to entry – Some of the UCS encoding schemes have the same labels as the UCS encoding form. However, they are used in different contexts. UCS encoding forms refer to in-memory and application interface representation of textual data. UCS encoding schemes refer to octet-serialized textual data.

3.25
extended collection
collection for which the entities can also consist of sequences of code points that are in Normalization Form C (NFC)

Note 1 to entry – Some collections such as 3 LATIN EXTENDED-A, 4 LATIN EXTENDED-B, 15 ARABIC EXTENDED, and many more, have the term 'extended' in their name. This does not make them extended collections.

Note 2 to entry – See Clause 21 for discussion of Normalization Form C.

Note 3 to entry – The sequences of code points are typically referenced by Named UCS Sequence Identifiers (NUSI) (see Clause 27).

3.26
fixed collection
collection in which every code point within the identified range(s) has a character allocated to it, and which is intended to remain unchanged in future editions of this International Standard

3.27
format character
character whose primary function is to affect the layout or processing of characters around it

Note 1 to entry – A format character generally does not have a visible representation of its own.

3.28
General Category
GC
value assigned to each UCS code point which determines its major class, such as letter, punctuation, and symbol

Note 1 to entry – Each value is defined as General Category property using a two-letter abbreviation in the Unicode Standard (see reference to the current Unicode Standard General Category in 2).

Note 2 to entry – When referred as a group containing all GC values sharing the same first letter, the group may be described using the first letter only. For example, 'L' stands for all letters 'Lu', 'Ll', 'Lt', 'Lm', and 'Lo'.

3.29
graphic character
character, other than a control function or a format character, that has a visual representation normally hand-written, printed, or displayed

3.30
graphic symbol
visual representation of a graphic character or of a composite sequence

3.31
high-surrogate code point
code point in the range D800 to DBFF reserved for the use of UTF-16

3.32
high-surrogate code unit
16-bit code unit in the range D800 to DBFF used in UTF-16 as the leading code unit of a surrogate pair (see 9.3)

3.33
ill-formed code unit sequence
UCS code unit sequence that purports to be in a UCS encoding form which does not conform to the specification of that encoding form

EXAMPLE – An unpaired surrogate code unit is an ill-formed code unit sequence.

3.34
ill-formed code unit sequence subset
non-empty subset of a code unit sequence X which does not contain any code unit which also belong to any minimal well-formed code unit sequence subset of X

Note 1 to entry – An ill-formed code unit sequence subset cannot overlap with a minimal well-formed code unit sequence.

3.35
interchange
transfer of character coded data from one user to another, using telecommunication means or interchangeable media

Note 1 to entry – Interchange implies data serialization and the use of a UCS encoding scheme.

3.36
interworking
process of permitting two or more systems, each employing different coded character sets, to meaningfully interchange character coded data

Note 1 to entry – Conversion between the two codes might be involved.

3.37
ISO/IEC 10646-1
former subdivision of ISO/IEC 10646 containing the specification of the overall architecture and the Basic Multilingual Plane (BMP)

Note 1 to entry – It is also referred to as Part 1 of ISO/IEC 10646.

Note 2 to entry – There are a first and a second Edition of ISO/IEC 10646-1.

3.38
ISO/IEC 10646-2
former subdivision of ISO/IEC 10646 containing the specification of the Supplementary Multilingual Plane (SMP), the Supplementary Ideographic Plane (SIP) and the Supplementary Special-purpose Plane (SSP)

Note 1 to entry – It is also referred to as Part 2 of ISO/IEC 10646.

Note 2 to entry – There is only a first edition of ISO/IEC 10646-2.

3.39
low-surrogate code point
code point in the range DC00 to DFFF reserved for the use of UTF-16

3.40
low-surrogate code unit
16-bit code unit in the range DC00 to DFFF used in UTF-16 as the trailing code unit of a surrogate pair (see 9.3)

3.41
minimal well-formed code unit sequence
well-formed code unit sequence that maps to a single UCS scalar value

3.42
mirrored character
character whose image is mirrored horizontally in text that is laid out from right to left

3.43
octet
8-bit code unit

Note 1 to entry – The value is expressed in hexadecimal notation from 00 to FF in this International Standard (see Annex K).

3.44
plane
subdivision of the UCS codespace consisting of contiguous 65 536 code points beginning at a multiple of 65 536 which can be identified by a number from 00 to 10

Note 1 to entry – The UCS codespace contain 17 planes.

3.45
presentation
process of writing, printing, or displaying a graphic symbol

3.46
presentation form
(in the presentation of some scripts) form of a graphic symbol representing a character that depends on the position of the character relative to other characters

3.47
private use plane
plane within this coded character set, the content of which is not specified in this International Standard

Note 1 to entry – Planes 0F and 10 are private use planes.

3.48
repertoire
specified set of characters that are represented in a coded character set

3.49
row
subdivision of a plane consisting of contiguous 256 code points beginning at a multiple of 256 which can be identified by a number from 00 to FF

3.50
script
set of graphic characters used for the written form of one or more languages

3.51
supplementary plane
plane other than Plane 00 of the UCS codespace

Note 1 to entry – A supplementary plane accommodates characters which have not been allocated to the Basic Multilingual Plane.

3.52
Supplementary Multilingual Plane for scripts and symbols
SMP
plane 01 of the UCS codespace

3.53
Supplementary Ideographic Plane
SIP
plane 02 of the UCS codespace

3.54
Supplementary Special-purpose Plane
SSP
plane 0E of the UCS codespace

3.55
surrogate pair
representation for a single character that consists of a sequence of two 16-bit code units, where the first value of the pair is a high-surrogate code unit and the second value is a low-surrogate code unit

3.56
Tertiary Ideographic Plane
TIP
plane 03 of the UCS codespace

3.57
UCS codespace
codespace consisting of the integers from 0 to 10FFFF (hexadecimal) available for assigning the repertoire of the UCS characters

3.58
UCS scalar value
any UCS code point except high-surrogate and low-surrogate code points