



SLOVENSKI STANDARD SIST ISO/IEC 10646:2013

01-december-2013

Nadomešča:

SIST ISO/IEC 10646:2008

Informacijska tehnologija - Univerzalni večoktetni nabor znakov (UCS)

Information technology -- Universal Coded Character Set (UCS)

iTeh STANDARD PREVIEW
Technologies de l'information -- Jeu universel de caractères codés (JUC)
(standards.iteh.ai)

Ta slovenski standard je istoveten z: ISO/IEC 10646:2012

<https://standards.iteh.ai/catalog/standards/sist/82f65aaf-cf7e-469c-8006-15039feb0ff/sist-iso-iec-10646-2013>

ICS:

35.040	Nabori znakov in kodiranje informacij	Character sets and information coding
--------	---------------------------------------	---------------------------------------

SIST ISO/IEC 10646:2013

en,fr,de

iTeh STANDARD PREVIEW
(standards.iteh.ai)

[SIST ISO/IEC 10646:2013](#)

<https://standards.iteh.ai/catalog/standards/sist/82f65aaf-cf7e-469c-8006-15039ffc9ff/sist-iso-iec-10646-2013>

INTERNATIONAL
STANDARD

ISO/IEC
10646

Third edition
2012-06-01

**Information technology — Universal
Coded Character Set (UCS)**

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

iTeh STANDARD PREVIEW
(standards.iteh.ai)

[SIST ISO/IEC 10646:2013](https://standards.iteh.ai/catalog/standards/sist/82f65aaf-cf7e-469c-8006-15039ffc99ff/sist-iso-iec-10646-2013)

<https://standards.iteh.ai/catalog/standards/sist/82f65aaf-cf7e-469c-8006-15039ffc99ff/sist-iso-iec-10646-2013>

Reference number
ISO/IEC 10646:2012(E)



iTeh STANDARD PREVIEW (standards.iteh.ai)

[SIST ISO/IEC 10646:2013](https://standards.iteh.ai/catalog/standards/sist/82f65aaf-cf7e-469c-8006-15039ffc9ff/sist-iso-iec-10646-2013)

<https://standards.iteh.ai/catalog/standards/sist/82f65aaf-cf7e-469c-8006-15039ffc9ff/sist-iso-iec-10646-2013>



COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2012

All rights reserved. Unless otherwise specified, no part of this publication may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying and microfilm, without permission in writing from either ISO at the address below or ISO's member body in the country of the requester.

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

CONTENTS

Foreword.....	vii
Introduction.....	viii
1 Scope	1
2 Conformance.....	1
2.1 General	1
2.2 Conformance of information interchange.....	2
2.3 Conformance of devices	2
3 Normative references	2
4 Terms and definitions	3
5 General structure of the UCS.....	9
6 Basic structure and nomenclature.....	9
6.1 Structure.....	9
6.2 Coding of characters.....	11
6.3 Types of code points	11
6.4 Naming of characters.....	12
6.5 Short identifiers for code points (UIDs).....	12
6.6 UCS Sequence Identifiers.....	13
6.7 Octet sequence identifiers.....	13
7 Revision and updating of the UCS.....	14
8 Subsets.....	14
8.1 Limited subset.....	14
8.2 Selected subset.....	14
9 UCS encoding forms	14
9.1 UTF-8.....	14
9.2 UTF-16.....	15
9.3 UTF-32 (UCS-4).....	16
10 UCS Encoding schemes	16
10.1 UTF-8.....	16
10.2 UTF-16BE	16
10.3 UTF-16LE.....	16
10.4 UTF-16	16
10.5 UTF-32BE	17
10.6 UTF-32LE.....	17
10.7 UTF-32.....	17
11 Use of control functions with the UCS.....	17
12 Declaration of identification of features	18
12.1 Purpose and context of identification	18
12.2 Identification of a UCS encoding form	19
12.3 Identification of subsets of graphic characters.....	19
12.4 Identification of control function set.....	19
12.5 Identification of the coding system of ISO/IEC 2022	20
13 Structure of the code charts and lists.....	20

ISO/IEC 10646:2012 (E)

14	Block and collection names	21
14.1	Block names	21
14.2	Collection names	21
15	Mirrored characters in bidirectional context	21
15.1	Mirrored characters	21
15.2	Directionality of bidirectional text	21
16	Special characters	22
16.1	Space characters	22
16.2	Currency symbols	22
16.3	Format Characters	22
16.4	Ideographic description characters	23
16.5	Variation selectors and variation sequences	23
17	Presentation forms of characters	26
18	Compatibility characters	26
19	Order of characters	26
20	Combining characters	27
20.1	Order of combining characters	27
20.2	Combining class and canonical ordering	27
20.3	Appearance in code charts	27
20.4	Alternate coded representations	27
20.5	Multiple combining characters	27
20.6	Collections containing combining characters	28
20.7	Combining Grapheme Joiner	28
21	Normalization forms	29
22	Special features of individual scripts and symbol repertoires	29
22.1	Hangul syllable composition method	29
22.2	Features of scripts used in India and some other South Asian countries	29
22.3	Byzantine musical symbols	30
22.4	Source references for pictographic symbols	30
23	Source references for CJK Ideographs	31
23.1	List of source references	31
23.2	Source references for CJK Unified Ideographs	33
23.3	Source reference presentation for CJK Unified Ideographs	34
23.4	Source references for CJK Compatibility Ideographs	36
23.5	Source references presentation for CJK Compatibility Ideographs	37
24	Character names and annotations	38
24.1	Entity names	38
24.2	Name formation	38
24.3	Single name	38
24.4	Name immutability	39
24.5	Name uniqueness	39
24.6	Character names for CJK Ideographs	40
24.7	Character names for Hangul syllables	40
25	Named UCS Sequence Identifiers	41

26	Structure of the Basic Multilingual Plane	43
27	Structure of the Supplementary Multilingual Plane for scripts and symbols (SMP)	45
28	Structure of the Supplementary Ideographic Plane (SIP)	47
29	Structure of the Tertiary Ideographic Plane (TIP)	47
30	Structure of the Supplementary Special-purpose Plane (SSP)	47
31	Code charts and lists of character names	48
31.1	Code chart	48
31.2	Character names list	48
31.3	Pointers to code charts and lists of character names	49
Annex A	(normative) Collections of graphic characters for subsets	2181
A.1	Collections of coded graphic characters	2181
A.2	Blocks lists	2186
A.3	Fixed collections of the whole UCS (except Unicode collections)	2188
A.4	CJK collections	2192
A.5	Other collections	2193
A.6	Unicode collections	2196
Annex B	(normative) List of combining characters	2207
Annex C	(normative) Transformation format for planes 01 to 10 of the UCS (UTF-16)	2208
Annex D	(normative) UCS Transformation Format 8 (UTF-8)	2209
Annex E	(normative) Mirrored characters in bidirectional context	2210
Annex F	(informative) Format characters	2211
F.1	General format characters	2211
F.2	Script-specific format characters	2213
F.3	Interlinear annotation characters	2214
F.4	Subtending format characters	2214
F.5	Contiguity operators	2214
F.6	Western musical symbols	2215
F.7	Language tagging using Tag characters	2216
Annex G	(informative) Alphabetically sorted list of character names	2218
Annex H	(informative) The use of "signatures" to identify UCS	2219
Annex I	(informative) Ideographic description characters	2220
Annex J	(informative) Recommendation for combined receiving/originating devices with internal storage	2223
Annex K	(informative) Notations of octet value representations	2224
Annex L	(informative) Character naming guidelines	2225
Annex M	(informative) Sources of characters	2228
Annex N	(informative) External references to character repertoires	2245
N.1	Methods of reference to character repertoires and their coding	2245
N.2	Identification of ASN.1 character abstract syntaxes	2245
N.3	Identification of ASN.1 character transfer syntaxes	2246
Annex P	(informative) Additional information on CJK Unified Ideographs	2247
Annex Q	(informative) Code mapping table for Hangul syllables	2248
Annex R	(informative) Names of Hangul syllables	2249

ISO/IEC 10646:2012 (E)

Annex S (informative) Procedure for the unification and arrangement of CJK Ideographs	2250
S.1 Unification procedure	2250
S.2 Arrangement procedure	2254
S.3 Source code separation examples.....	2254
S.4 Non-unification examples.....	2259
Annex T (informative) Language tagging using Tag Characters.....	2261
Annex U (informative) Characters in identifiers.....	2262

iTeh STANDARD PREVIEW **(standards.iteh.ai)**

[SIST ISO/IEC 10646:2013](https://standards.iteh.ai/catalog/standards/sist/82f65aaf-cf7e-469c-8006-15039ffc99ff/sist-iso-iec-10646-2013)

<https://standards.iteh.ai/catalog/standards/sist/82f65aaf-cf7e-469c-8006-15039ffc99ff/sist-iso-iec-10646-2013>

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 JTC1.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of the joint technical committee is to prepare International Standards. Draft International Standards adopted by the joint technical committee are circulated to national bodies for voting. Publication as an International Standard requires approval by at least 75% of the national bodies casting a vote.

Attention is drawn to the possibility that some of the elements of ISO/IEC 10646 may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

This International Standard was prepared by Joint Technical Committee ISO/IEC JTC1, Subcommittee SC 2, on behalf of ISO and IEC.

This third edition cancels and replaces ISO/IEC 10646:2011, which has been amended by ISO/IEC 10646:2011/Amendment 1 and ISO/IEC 10646:2011/Amendment 2.

iteh STANDARD PREVIEW
(standards.iteh.ai)

[SIST ISO/IEC 10646:2013](https://standards.iteh.ai/catalog/standards/sist/82f65aaf-cf7e-469c-8006-15039ffc99ff/sist-iso-iec-10646-2013)

<https://standards.iteh.ai/catalog/standards/sist/82f65aaf-cf7e-469c-8006-15039ffc99ff/sist-iso-iec-10646-2013>

ISO/IEC 10646:2012 (E)**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 109 000 characters from the world's scripts.

This International Standard contains material which may only be available to users who obtain their copy in a machine readable format. That material consists of the following printable files:

- EmojiSrc.txt
- CJKU_SR.txt
- CJKC_SR.txt
- NUSI.txt
- IICORE.txt
- JIEx.txt
- Allnames.txt
- HanguSy.txt.

iTeh STANDARD PREVIEW
(standards.iteh.ai)

[SIST ISO/IEC 10646:2013](https://standards.iteh.ai/catalog/standards/sist/82f65aaf-cf7e-469c-8006-15039ffc99ff/sist-iso-iec-10646-2013)

<https://standards.iteh.ai/catalog/standards/sist/82f65aaf-cf7e-469c-8006-15039ffc99ff/sist-iso-iec-10646-2013>

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.

NOTE – The Unicode Standard, Version 6.1 includes a set of characters, names, and coded representations that are identical with those in this International Standard. It additionally provides details of character properties, processing algorithms, and definitions that are useful to implementers.

2 Conformance

2.1 General

Whenever private use characters are used as specified in this International Standard, the characters themselves shall not be covered by these conformance requirements.

ISO/IEC 10646:2012 (E)

2.2 Conformance of information interchange

A code unit sequence (CC-data-element) within coded information for interchange is in conformance with this International Standard if

- a) all the coded representations of graphic characters within that code unit sequence conform to clause 6, to an identified encoding form chosen from clause 9, and to an identified encoding scheme chosen from clause 10;
- b) all the graphic characters represented within that code unit sequence are taken from those within an identified subset (see 8);
- c) all the coded representations of control functions within that code unit sequence conform to clause 11.

A claim of conformance shall identify the adopted encoding form, the adopted encoding scheme, and the adopted subset by means of a list of collections and/or characters.

2.3 Conformance of devices

A device is in conformance with this International Standard if it conforms to the requirements of item a) below, and either or both of items b) and c).

A claim of conformance shall identify the document that contains the description specified in a) below, and shall identify the adopted encoding form(s), the adopted encoding scheme(s), and the adopted subset (by means of a list of collections and/or characters), and the selection of control functions adopted in accordance with clause 11.

- a) **Device description:** A device that conforms to this International Standard shall be the subject of a description that identifies the means by which the user may supply characters to the device and/or may recognize them when they are made available to the user, as specified respectively, in subclauses b) and c) below.
- b) **Originating device:** An originating device shall allow its user to supply any characters from an adopted subset, and be capable of transmitting their coded representations within a code unit sequence in accordance with the adopted encoding form and adopted encoding scheme. As such, the originating device shall not emit ill-formed code unit sequences.
- c) **Receiving device:** A receiving device shall be capable of receiving and interpreting any coded representation of characters that are within a code unit sequence in accordance with the adopted encoding form and the adopted encoding scheme, and shall make any corresponding characters from the adopted subset available to the user in such a way that the user can identify them. The receiving device shall treat ill-formed code unit sequences as an error condition and shall not interpret such data as character sequences.

Any corresponding characters that are not within the adopted subset shall be indicated to the user. The way used for indicating them need not distinguish them from each other.

NOTE 1 – The manner in which a user is notified of either an error condition or characters not within the adopted subset is not specified by this International Standard.

NOTE 2 – See also Annex J for receiving devices with retransmission capability.

3 Normative references

The following referenced documents are indispensable for the application 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/tr9-25.html>.

Unicode Standard Annex, UAX #15, *Unicode Normalization Forms*:
<http://www.unicode.org/reports/tr15/tr15-35.html>.

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

Unicode Standard Version 6.1, *Chapter 4, Character Properties*
<http://www.unicode.org/versions/Unicode6.1.0/ch04.pdf>
Section 4.3, Combining Classes – Normative
Section 4.5, General Category – Normative
Section 4.7, Bidi Mirrored – Normative

4 Terms and definitions

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

4.1

base character

graphic character which is not a combining character

NOTE 1 – 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 – A base character typically does not graphically combine with preceding characters. They can be exceptions for some complex writing systems.

4.2

Basic Multilingual Plane BMP

plane 00 of the UCS codespace

ITeH STANDARD PREVIEW
(standards.iteh.ai)

4.3

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

SIST ISO/IEC 10646:2013

<http://standards.iteh.ai/catalog/standards/sist/10646-2013/iso-iec-10646-2013>

4.4

canonical form

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

NOTE – 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 4.23).

4.5

character

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

NOTE – A character can be represented by a sequence of one or several coded characters.

4.6

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

4.7

code chart

code table

rectangular array showing the representation of coded characters allocated within a range of the UCS codespace

ISO/IEC 10646:2012 (E)**4.8****coded character**

association between a character and a code point

4.9**coded character set**

set of coded characters

4.10**code point****code position**

value in the UCS codespace

4.11**code unit**

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

NOTE – 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.

4.12**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 – Such sequence can contain code units associated with any type of code points.

NOTE 2 – Unlike previous editions of ISO/IEC 10646, this International Standard no longer uses 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.

4.13**collection**

numbered and named set of entities

NOTE 1 – 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 4.25 for extended collection).

NOTE 2 – 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.

4.14**combining character**

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

NOTE – 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 4.17).

4.15**combining class**

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

4.16**compatibility character**

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

4.17**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 4.14)

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

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

4.18**control character**

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

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

4.19**control function**

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

4.20**decomposition mapping**

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

4.21**default state**

state that is assumed when no state has been explicitly specified

NOTE – See F.2.1 and F.2.3.

4.22**device**

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

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

4.23**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 – This International Standard specifies UTF-8, UTF-16, and UTF-32.

4.24**encoding scheme**

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

NOTE – 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.

4.25**extended collection**

collection for which the entities can also consist of sequences of code points that are in normalization form NFC

NOTE 1 – 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.