

INFOTEHNOLOOGIA
Universaalne koodimärgistik (UCS)

Information technology
Universal Coded Character Set (UCS)
(ISO/IEC 10646:2014)

EESTI STANDARDI EESSÕNA**NATIONAL FOREWORD**

| | |
|---|---|
| See Eesti standard EVS-ISO/IEC 10646:2014 „Infotehnoloogia. Universaalne koodimärgistik (UCS)“ sisaldab rahvusvahelise standardi ISO/IEC 10646:2014 „Information technology – Universal Coded Character Set (UCS)“ identset ingliskeelset teksti. | This Estonian Standard EVS-ISO/IEC 10646:2014 consists of the identical English text of the International Standard ISO/IEC 10646:2014 „Information technology – Universal Coded Character Set (UCS)“. |
| Ettepaneku rahvusvahelise standardi ümbertrüki meetodil ülevõtuks on esitanud EVS/TK 4, standardi avaldamist on korraldanud Eesti Standardikeskus. | Proposal to adopt the International Standard by reprint method has been presented by EVS/TK 4, the Estonian standard has been published by the Estonian Centre for Standardisation. |
| Standard EVS-ISO/IEC 10646:2014 on jõustunud sellekohase teate avaldamisega EVS Teataja 2015. aasta jaanuarikuu numbris. | This standard has been endorsed with a notification published in the official bulletin of the Estonian Centre for Standardisation. |
| Standard on kättesaadav Eesti Standardikeskusest. | The standard is available from the Estonian Centre for Standardisation. |

Käsitlusala

See rahvusvaheline standard kirjeldab universaalset koodimärgistikku (UCS). See on rakendatav maailma keelte ja lisasümbolite esituseks, edastamiseks, vahetamiseks, töötlemiseks, talletamiseks, sisestamiseks ja esitamiseks kirjalikus vormis.

See rahvusvaheline standard:

- täpsustab selle rahvusvahelise standardi arhitektuuri;
- defineerib selles rahvusvahelises standardis kasutatud termineid;
- kirjeldab koodimärgistiku koodiruumi üldstruktuuri;
- kirjeldab UCS-i mitmekeelset põhitasandit (BMP);
- kirjeldab UCS-i lisatasandeid: mitmekeelne lisatasand (SMP), ideograafiline lisatasand (SIP), tertsiaarne lisatasand (TIP) ja eriotstarbeline lisatasand (SSP);
- määratleb graafiliste märkide kogumi, mida kasutatakse ülemaailmselt skriptides ja loomulike keelte kirja pildis;
- täpsustab graafiliste märkide ja vormingumärkide nimesid BMP, SMP, SIP, TIP, SSP ning nende kodeeritud esituste jaoks UCS-koodiruumis;
- täpsustab juhtmärkide ja privaاتمärkide kodeeritud esitust;
- täpsustab kolme UCS-i kodeerimisvormi: UTF-8, UTF-16 ja UTF-32;
- täpsustab seitse UCS-i kodeerimisskeemi: UTF-8, UTF-16, UTF-16BE, UTF-16LE, UTF-32, UTF-32BE ja UTF-32LE;
- täpsustab selle koodimärgistiku tulevaste lisandite haldust.

UCS on standardis ISO/IEC 2022 kirjeldatust erinev kodeerimissüsteem. Meetod, kuidas eristada UCS-i standardist ISO/IEC 2022, on täpsustatud jaotises 12.2.

Graafilisele märgile omistatakse standardis ainult üks märgi koodipositsioon, mis asub kas BMP-s või mõnel lisatasandil.

Tagasisidet standardi sisu kohta on võimalik edastada, kasutades EVS-i veebilehel asuvat tagasiside vormi või saates e-kirja meiliaadressile standardiosakond@evs.ee.

ICS 35.040

Standardite reprodutseerimise ja levitamise õigus kuulub Eesti Standardikeskusele

Andmete paljundamine, taastekitamine, kopeerimine, salvestamine elektroonsesse süsteemi või edastamine ükskõik millises vormis või millisel teel ilma Eesti Standardikeskuse kirjaliku loata on keelatud.

Kui Teil on küsimusi standardite autorikaitse kohta, võtke palun ühendust Eesti Standardikeskusega:

Aru 10, 10317 Tallinn, Eesti; www.evs.ee; telefon 605 5050; e-post info@evs.ee

The right to reproduce and distribute standards belongs to the Estonian Centre for Standardisation

No part of this publication may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, without a written permission from the Estonian Centre for Standardisation.

If you have any questions about copyright, please contact Estonian Centre for Standardisation:

Aru 10, 10317 Tallinn, Estonia; www.evs.ee; phone 605 5050; e-mail info@evs.ee

CONTENTS

| | |
|--|------|
| Foreword..... | vii |
| Introduction | viii |
| 1 Scope | 1 |
| 2 Conformance | 1 |
| 2.1 General | 1 |
| 2.2 Conformance of information interchange | 1 |
| 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 General | 14 |
| 8.2 Limited subset | 14 |
| 8.3 Selected subset | 14 |
| 9 UCS encoding forms | 14 |
| 9.1 General | 14 |
| 9.2 UTF-8 | 14 |
| 9.3 UTF-16 | 15 |
| 9.4 UTF-32 (UCS-4) | 16 |
| 10 UCS Encoding schemes | 16 |
| 10.1 General | 16 |
| 10.2 UTF-8 | 16 |
| 10.3 UTF-16BE | 16 |
| 10.4 UTF-16LE | 16 |
| 10.5 UTF-16 | 16 |
| 10.6 UTF-32BE | 17 |
| 10.7 UTF-32LE | 17 |
| 10.8 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 scheme | 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 |
| 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 | General | 22 |
| 16.2 | Space characters | 22 |
| 16.3 | Currency symbols | 22 |
| 16.4 | Format characters..... | 22 |
| 16.5 | Ideographic description characters | 23 |
| 16.6 | Variation selectors and variation sequences | 23 |
| 17 | Presentation forms of characters | 24 |
| 18 | Compatibility characters | 25 |
| 19 | Order of characters | 25 |
| 20 | Combining characters | 25 |
| 20.1 | Order of combining characters..... | 25 |
| 20.2 | Combining class and canonical ordering | 26 |
| 20.3 | Appearance in code charts | 26 |
| 20.4 | Alternate coded representations | 26 |
| 20.5 | Multiple combining characters | 26 |
| 20.6 | Collections containing combining characters..... | 27 |
| 20.7 | Combining Grapheme Joiner | 27 |
| 21 | Normalization forms | 27 |
| 22 | Special features of individual scripts and symbol repertoires | 28 |
| 22.1 | Hangul syllable composition method | 28 |
| 22.2 | Features of scripts used in India and some other South Asian countries..... | 28 |
| 22.3 | Byzantine musical symbols | 29 |
| 22.4 | Source references for pictographic symbols..... | 29 |
| 23 | Source references for CJK Ideographs | 29 |
| 23.1 | List of source references..... | 29 |
| 23.2 | Source references file for CJK Ideographs | 32 |
| 23.3 | Source reference presentation for CJK Unified Ideographs | 34 |
| 23.4 | Source references presentation for CJK Compatibility Ideographs | 36 |
| 24 | Character names and annotations | 37 |
| 24.1 | Entity names | 37 |
| 24.2 | Name formation..... | 37 |
| 24.3 | Single name | 38 |
| 24.4 | Name immutability..... | 38 |
| 24.5 | Name uniqueness | 38 |

| | | |
|---------|---|------|
| 24.6 | Character names for CJK Ideographs | 39 |
| 24.7 | Character names for Hangul syllables | 39 |
| 25 | Named UCS Sequence Identifiers | 41 |
| 26 | Structure of the Basic Multilingual Plane | 42 |
| 27 | Structure of the Supplementary Multilingual Plane for scripts and symbols (SMP) | 44 |
| 28 | Structure of the Supplementary Ideographic Plane (SIP) | 46 |
| 29 | Structure of the Tertiary Ideographic Plane (TIP) | 46 |
| 30 | Structure of the Supplementary Special-purpose Plane (SSP) | 46 |
| 31 | Code charts and lists of character names | 47 |
| 31.1 | General | 47 |
| 31.2 | Code chart | 47 |
| 31.3 | Character names list | 47 |
| 31.4 | Summary of standardized variation sequences | 48 |
| 31.5 | Pointers to code charts and lists of character names | 48 |
| Annex A | (normative) Collections of graphic characters for subsets | 2381 |
| A.1 | Collections of coded graphic characters | 2381 |
| A.2 | Blocks lists | 2387 |
| A.3 | Fixed collections of the whole UCS (except Unicode collections) | 2389 |
| A.4 | CJK collections | 2393 |
| A.5 | Other collections | 2393 |
| A.6 | Unicode collections | 2397 |
| Annex B | (normative) List of combining characters | 2410 |
| Annex C | (normative) Transformation format for planes 01 to 10 of the UCS (UTF-16) | 2411 |
| Annex D | (normative) UCS Transformation Format 8 (UTF-8) | 2412 |
| Annex E | (normative) Mirrored characters in bidirectional context | 2413 |
| Annex F | (informative) Format characters | 2414 |
| F.1 | General format characters | 2414 |
| F.2 | Script-specific format characters | 2416 |
| F.3 | Interlinear annotation characters | 2417 |
| F.4 | Subtending format characters | 2417 |
| F.5 | Shorthand format characters | 2418 |
| F.6 | Invisible mathematical operators | 2418 |
| F.7 | Western musical symbols | 2418 |
| F.8 | Language tagging using Tag characters | 2419 |
| Annex G | (informative) Alphabetically sorted list of character names | 2421 |
| Annex H | (informative) The use of “signatures” to identify UCS | 2422 |
| Annex I | (informative) Ideographic description characters | 2423 |
| I.1 | General | 2423 |
| I.2 | Syntax of an ideographic description sequence | 2423 |
| I.3 | Individual definitions of the ideographic description characters | 2423 |
| Annex J | (informative) Recommendation for combined receiving/originating devices with internal storage | 2426 |
| Annex K | (informative) Notations of octet value representations | 2427 |

| | |
|---|------|
| Annex L (informative) Character naming guidelines | 2428 |
| Annex M (informative) Sources of characters | 2431 |
| Annex N (informative) External references to character repertoires | 2449 |
| N.1 Methods of reference to character repertoires and their coding | 2449 |
| N.2 Identification of ASN.1 character abstract syntaxes | 2449 |
| N.3 Identification of ASN.1 character transfer syntaxes | 2450 |
| Annex P (informative) Additional information on CJK Unified Ideographs | 2451 |
| Annex Q (informative) Code mapping table for Hangul syllables | 2454 |
| Annex R (informative) Names of Hangul syllables | 2455 |
| Annex S (informative) Procedure for the unification and arrangement of CJK Ideographs | 2456 |
| S.1 Unification procedure | 2456 |
| S.2 Arrangement procedure | 2459 |
| S.3 Source separation examples | 2460 |
| S.4 Non-unification examples | 2465 |
| Annex T (informative) Language tagging using Tag Characters | 2466 |
| Annex U (informative) Characters in identifiers | 2467 |

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 meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the WTO principles in the Technical Barriers to Trade (TBT) see the following URL: [Foreword - Supplementary information](#)

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

This fourth edition cancels and replaces the third edition (ISO/IEC 10646:2012), which has been technically revised. It also incorporates the Amendment ISO/IEC 10646:2012/Amd.1:2013.

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 120 000 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
- UCSVariants.txt
- CJKSrc.txt
- NUSI.txt
- JIExt.txt
- Allnames.txt
- HangulSy.txt.

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 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.

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 Clause 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 sub-clauses 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/tr9-31.html>.

Unicode Standard Annex, UAX #15, *Unicode Normalization Forms*:

<http://www.unicode.org/reports/tr15/tr15-41.html>.

Unicode Technical Standard, UTS #37, *Ideographic Variation Database*:

<http://www.unicode.org/reports/tr37/tr37-8.html>.

Unicode Standard Version 6.2, *Chapter 4, Character Properties*

<http://www.unicode.org/versions/Unicode7.0.0/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. There are exceptions for some complex writing systems.

4.2

Basic Multilingual Plane

BMP

plane 00 of the UCS codespace

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

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 graphic symbol 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

4.8

coded character

association between a character and a code point

4.9

coded character set

set of coded characters