INFOTEHNOLOOGIA
Universaalne koodimärgistik (UCS)

Information technology
Universal Coded Character Set (UCS)
(ISO/IEC 10646:2012)

EESTI STANDARDIKESKUS
ESTONIAN CENTRE FOR STANDARDISATION

EESTI STANDARDI EESSÕNA	NATIONAL FOREWORD
See Eesti standard EVS-ISO/IEC 10646:2012 "Infotehnoloogia. Universaalne koodimärgistik (UCS)" sisaldab rahvusvahelise standardi ISO/IEC 10646:2012 „Information technology – Universal Coded Character Set (UCS)" identset ingliskeelset teksti.	This Estonian Standard EVS-ISO 10646:2012 consists of the identical English text of the International Standard ISO/IEC 10646:2012 "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 10646:2012 on jõustunud sellekohase teate avaldamisega EVS Teataja 2012. aasta novembrikuu 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 UCSi mitmekeelset põhitasandit (BMP);
- kirjeldab UCSi 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 kirjapildis;
- täpsustab graafiliste märkide ja vormingumärkide nimesid BMP, SMP, SIP, TIP, SSP ning nende kodeeritud esituste jaoks UCS koodiruumis;
- täpsustab juhtmärkide ja privaatmärkide kodeeritud esitust;
- täpsustab kolme UCS kodeerimisvormi: UTF-8, UTF-16 ja UTF-32;
- täpsustab seitse UCS 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 UCSi standardist ISO/IEC 2022, on täpsustatud jaotises 12.2.

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

MÄRKUS Unicode standardi versioon 6.1 sisaldab märkide, nimede ja kodeeritud esituste kogumit, mis on selle rahvusvahelise standardi omaga identsed. Lisaks annab see üksikasjalikumat teavet märkide omaduste, töötlusalgoritmide ja definitsioonide kohta, mis on rakendajatele kasulikud.

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

CONTENTS

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
- 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.

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.

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

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 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