INFOTEHNOLOOGIA
Universaalne koodimärgistik (UCS)

Information technology
Universal Coded Character Set (UCS)
(ISO/IEC 10646:2017, identical
+ ISO/IEC 10646:2017/Amd 1:2019, identical
+ ISO/IEC 10646:2017/Amd 2:2019, identical)

EESTI STANDARDIKESKUS
ESTONIAN CENTRE FOR STANDARDISATION

EESTI STANDARDI EESSÕNA	NATIONAL FOREWORD
See Eesti standard EVS-ISO/IEC 10646:2020 „Infotehnoloogia. Universaalne koodimärgistik (UCS)“ sisaldab rahvusvahelise standardi ISO/IEC 10646:2017 „Information technology. Universal Coded Character Set (UCS)“ ning selle muudatuste ISO/IEC 10646:2017/Amd 1:2019 ja ISO/IEC 10646:2017/Amd 2:2019 identset ingliskeelset teksti.	This Estonian Standard EVS-ISO/IEC 10646:2020 consists of the identical English text of the International Standard ISO/IEC 10646:2017 „Information technology. Universal Coded Character Set (UCS)“ including its Amendments ISO/IEC 10646:2017/Amd 1:2019 and ISO/IEC 10646:2017/Amd 2:2019.
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 10646:2020 on jõustunud sellekohase teate avaldamisega EVS Teataja 2020. aasta veebruarikuu numbris.	Standard EVS-ISO/IEC 10646:2020 has been endorsed with a notification published in the February 2020 issue of the official bulletin of the Estonian Centre for Standardisation.
Standard on kättesaadav Eesti Standardikeskusest.	This standard is available from the Estonian Centre for Standardisation.

Käsitlusala

See rahvusvaheline standard kirjeldab universaalset koodimärgistikku (*Universal Coded Character Set*, 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 struktuuri;
- määratleb selles rahvusvahelises standardis kasutatud termineid;
- kirjeldab koodimärgistiku koodiruumi üldstruktuuri;
- kirjeldab UCS-i mitmekeelset põhitasandit (*Basic Multilingual Plane*, BMP);
- kirjeldab UCS-i lisatasandeid: mitmekeelne lisatasand (*Supplementary Multilingual Plane*, SMP), ideograafiline lisatasand (*Supplementary Ideographic Plane*, SIP), tertsiaarne lisatasand (*Tertiary Ideographic Plane*, TIP) ja eriotstarbeline lisatasand (*Supplementary Special-purpose Plane*, SSP);
- määratleb kirjamärkide kogumi, mida kasutatakse ülemaailmselt skriptides ja loomulike keelte kirjapildis;
- täpsustab kirjamärkide ja vormingumärkide nimesid BMP, SMP, SIP, TIP, SSP ning nende kodeeritud esituste jaoks UCS-koodiruumis;
- täpsustab juhtmärkide ja privaatmärkide kodeeritud esitust;
- täpsustab kolme UCS-i kodeerimisvormi: UTF-8, UTF-16 ja UTF-32;
- täpsustab seitset 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.

Kirjamärgile 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 35.040.10

Contents

FOREWORD ... 8
INTRODUCTION ... 9
1 SCOPE ... 10
2 NORMATIVE REFERENCES ... 11
3 TERMS AND DEFINITIONS ... 11
4 CONFORMANCE ... 19
4.1 General ... 19
4.2 Conformance of information interchange ... 20
4.3 Conformance of devices ... 20
5 GENERAL STRUCTURE OF THE UCS ... 20
6 BASIC STRUCTURE AND NOMENCLATURE ... 22
6.2 Coding of characters ... 22
6.3 Types of code points ... 23
6.3.2 Graphic characters ... 23
6.3.3 Format characters ... 24
6.3.4 Control characters ... 24
6.3.5 Private use characters ... 24
6.3.6 Surrogate code points ... 24
6.3.7 Noncharacter code points ... 24
6.3.8 Reserved code points ... 24
6.4 Naming of characters ... 24
6.5 Short identifiers for code points (UIDs) ... 25
6.6 UCS Sequence Identifiers ... 25
6.7 Octet sequence identifiers ... 26
7 REVISION AND UPDATING OF THE UCS ... 26
8 SUBSETS ... 26
8.1 General ... 26
8.2 Limited subset ... 27
8.3 Selected subset ... 27
9 UCS ENCODING FORMS ... 27
9.1 General ... 27
9.2 UTF-8 ... 27
9.3 UTF-16 ... 28
9.4 UTF-32 (UCS-4) ... 29
10 UCS ENCODING SCHEMES ... 29
10.1 General ... 29
10.2 UTF-8 ... 29
10.3 UTF-16BE ... 29
10.4 UTF-16LE ... 29
10.5 UTF-16 ... 30
10.6 UTF-32BE ... 30
10.7 UTF-32LE ... 30
10.8 UTF-32 ... 30
11 USE OF CONTROL FUNCTIONS WITH THE UCS ... 30
12 DECLARATION OF IDENTIFICATION OF FEATURES ... 33
12.1 Purpose and context of identification ... 33
12.2 Identification of a UCS encoding scheme ... 33
12.3 Identification of subsets of graphic characters ... 34

12.4 Identification of control function set 34
12.5 Identification of the coding system of ISO/IEC 2022 34
13 STRUCTURE OF THE CODE CHARTS AND LISTS 35
14 BLOCK AND COLLECTION NAMES 36
14.1 Block names 36
14.2 Collection names 36
15 MIRRORED CHARACTERS IN BIDIRECTIONAL CONTEXT 36
15.1 Mirrored characters 36
15.2 Directionality of bidirectional text 36
16 SPECIAL CHARACTERS 37
16.1 General 37
16.3 Currency symbols 37
16.4 Format characters 38
16.5 Ideographic description characters 40
16.6 Variation selectors and variation sequences 40
16.6.1 General 40
16.6.2 Standardized variation sequences 40
16.6.3 Ideographic variation sequences 42
17 PRESENTATION FORMS OF CHARACTERS 42
18 COMPATIBILITY CHARACTERS 42
19 ORDER OF CHARACTERS 43
20 COMBINING CHARACTERS 43
20.1 Order of combining characters 43
20.2 Combining class and canonical ordering 43
20.3 Appearance in code charts 43
20.4 Alternate coded representations 43
20.5 Multiple combining characters 44
20.6 Collections containing combining characters 45
20.7 Combining Grapheme Joiner 45
21 NORMALIZATION FORMS 45
22 SPECIAL FEATURES OF INDIVIDUAL SCRIPTS AND SYMBOL REPERTOIRES 46
22.1 Hangul syllable composition method 46
22.2 Features of scripts used in India and some other South Asian countries 46
22.3 Byzantine musical symbols 47
22.4 Source references for pictographic symbols 47
23 SOURCE REFERENCES FOR CJK IDEOGRAPHS 47
23.1 List of source references 47
23.2 Source references file for CJK ideographs 52
23.3 Source reference presentation for CJK Unified ideographs 55
23.3.1 General 55
23.3.2 Source reference presentation for CJK UNIFIED IDEOGRAPHS block 56
23.3.3 Source reference presentation for CJK UNIFIED IDEOGRAPHS EXTENSION A 56
23.3.4 Source reference presentation for CJK UNIFIED IDEOGRAPHS EXTENSION B 57
23.3.5 Source reference presentation for CJK UNIFIED IDEOGRAPHS EXTENSION C, D, E, and F 57
23.4 Source references presentation for CJK Compatibility ideographs 58
24 SOURCE REFERENCES FOR TANGUT IDEOGRAPHS 58
24.1 List of source references 58
24.2 Source reference file for Tangut ideographs 59
24.3 Source reference presentation for Tanguts ideographs 60
25 SOURCE REFERENCES FOR NÜSHU CHARACTERS 60

25.1 List of source references 60
25.2 Source reference file for Nüshu characters 61
26 CHARACTER NAMES AND ANNOTATIONS 62
26.1 Entity names 62
26.2 Name formation 62
26.3 Single name 63
26.4 Name immutability 63
26.5 Name uniqueness 63
26.5.1 Block names 63
26.5.2 Collection names 63
26.5.3 Character names, character name aliases, and named UCS sequence identifiers 63
26.5.4 Determining uniqueness 63
26.6 Character names for CJK ideographs 64
26.7 Character names for Tangut ideographs 64
26.8 Character names for Nüshu characters 64
26.9 Character names for Hangul syllables 64
27 NAMED UCS SEQUENCE IDENTIFIERS 66
28 STRUCTURE OF THE BASIC MULTILINGUAL PLANE 68
29 STRUCTURE OF THE SUPPLEMENTARY MULTILINGUAL PLANE FOR SCRIPTS AND SYMBOLS (SMP) 70
30 STRUCTURE OF THE SUPPLEMENTARY IDEOGRAPHIC PLANE (SIP) 72
31 STRUCTURE OF THE TERTIARY IDEOGRAPHIC PLANE (TIP) 73
32 STRUCTURE OF THE SUPPLEMENTARY SPECIAL-PURPOSE PLANE (SSP) 73
33 CODE CHARTS AND LISTS OF CHARACTER NAMES 74
33.1 General 74
33.2 Code chart 74
33.3 Character names list 74
33.4 Summary of standardized variation sequences 76
33.5 Code charts and lists of character names 76
Annex A (normative) Collections of graphic characters for subsets 2632
Annex B (normative) List of combining characters 2673
Annex C (normative) Transformation format for planes 01 to 10 of the UCS (UTF-16) 2674
Annex D (normative) UCS Transformation Format 8 (UTF-8) 2675
Annex E (normative) Mirrored characters in bidirectional context 2676
Annex F (informative) Format characters 2677
Annex G (informative) Alphabetically sorted list of character names 2685
Annex H (informative) The use of “signatures” to identify UCS 2686
Annex I (informative) Ideographic description characters 2687
Annex J (informative) Recommendation for combined receiving/originating devices with internal storage 2690
Annex K (informative) Notations of octet value representations 2691
Annex L (informative) Character naming guidelines 2692
Annex M (informative) Sources of characters 2696
Annex N (informative) External references to character repertoires 2721
Annex P (informative) Additional information on CJK Unified ideographs 2724

Annex Q (informative) Code mapping table for Hangul syllables....2728
Annex R (informative) Names of Hangul syllables....2729
Annex S (informative) Procedure for the unification and arrangement of CJK ideographs....2730
Annex T (informative) Language tagging using Tag Characters....2742
Annex U (informative) Characters in identifiers....2743

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.

INTERNATIONAL STANDARD ISO/IEC 10646:2017

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