

Information technology — Universal coded character set (UCS)

National foreword

This British Standard is the UK implementation of ISO/IEC 10646:2020. It supersedes BS ISO/IEC 10646:2017, which is withdrawn.

The UK participation in its preparation was entrusted to Technical Committee IST/5, Programming languages, their environments and system software interfaces.

A list of organizations represented on this committee can be obtained on request to its committee manager.

This publication does not purport to include all the necessary provisions of a contract. Users are responsible for its correct application.

© The British Standards Institution 2021
Published by BSI Standards Limited 2021

ISBN 978 0 539 02565 1

ICS 35.040.10

Compliance with a British Standard cannot confer immunity from legal obligations.

This British Standard was published under the authority of the Standards Policy and Strategy Committee on 31 January 2021.

Amendments/corrigenda issued since publication

Date	Text affected
------	---------------

INTERNATIONAL STANDARD

ISO/IEC
10646

Sixth edition
2020-12

Information technology — Universal coded character set (UCS)

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



Reference number
ISO/IEC 10646:2020(E)



COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2020

All rights reserved. Unless otherwise specified, or required in the context of its implementation, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier; Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

ISO/IEC 10646:2020 (E)

CONTENTS

1	Scope.....	1
2	Normative references	1
3	Terms and definitions.....	2
4	Conformance	8
4.1	General	8
4.2	Conformance of information interchange	8
4.3	Conformance of devices.....	8
5	Electronic data attachments	9
6	General structure of the UCS.....	10
7	Basic structure and nomenclature	11
7.1	Structure.....	11
7.2	Coding of characters.....	12
7.3	Types of code points.....	12
7.4	Naming of characters	13
7.5	Short identifiers for code points (UIDs)	14
7.6	UCS Sequence Identifiers.....	14
7.7	Octet sequence identifiers	15
8	Revision and updating of the UCS	15
9	Subsets.....	15
9.1	General	15
9.2	Limited subset	15
9.3	Selected subset.....	15
10	UCS encoding forms.....	15
10.1	General	15
10.2	UTF-8.....	15
10.3	UTF-16.....	16
10.4	UTF-32.....	17
11	UCS encoding schemes	17
11.1	General	17
11.2	UTF-8.....	17
11.3	UTF-16BE	17
11.4	UTF-16LE.....	18
11.5	UTF-16.....	18
11.6	UTF-32BE	18
11.7	UTF-32LE.....	18
11.8	UTF-32.....	19
12	Use of control functions with the UCS	19
13	Declaration of identification of features	20
13.1	Purpose and context of identification	20
13.2	Identification of a UCS encoding scheme.....	20
13.3	Identification of subsets of graphic characters.....	21

ISO/IEC 10646:2020 (E)

13.4	Identification of control function set.....	21
13.5	Identification of the coding system of ISO/IEC 2022	21
14	Structure of the code charts and lists	22
15	Block and collection names.....	22
15.1	Block names.....	22
15.2	Collection names	23
16	Mirrored characters in bidirectional context.....	23
16.1	Mirrored characters	23
16.2	Directionality of bidirectional text.....	23
17	Special characters.....	23
17.1	General	23
17.2	Space characters	23
17.3	Currency symbols	24
17.4	Format characters.....	24
17.5	Ideographic description characters	24
17.6	Variation selectors and variation sequences	25
18	Presentation forms of characters	27
19	Compatibility characters.....	27
20	Order of characters.....	27
21	Combining characters.....	28
21.1	Order of combining characters	28
21.2	Combining class and canonical ordering.....	28
21.3	Appearance in code charts	28
21.4	Alternate coded representations	28
21.5	Multiple combining characters	28
21.6	Collections containing combining characters	29
21.7	Combining Grapheme Joiner.....	29
22	Normalization forms.....	29
23	Special features of individual scripts and symbol repertoires.....	30
23.1	Hangul syllable composition method.....	30
23.2	Features of scripts used in India and some other South Asian countries.....	30
23.3	Byzantine musical symbols	31
23.4	Source references for pictographic symbols	31
24	Source references for CJK ideographs	32
24.1	List of source references.....	32
24.2	Source references file for CJK ideographs	35
24.3	Source reference presentation for CJK Unified ideographs.....	37
24.4	Source references presentation for CJK Compatibility ideographs	40
25	Source references for Tangut ideographs.....	40
25.1	List of source references.....	40
25.2	Source reference file for Tangut ideographs	41
25.3	Source reference presentation for Tanguts ideographs.....	42
26	Source references for Nüshu characters.....	42

ISO/IEC 10646:2020 (E)

26.1	List of source references.....	42
26.2	Source reference file for Nüshu characters	42
27	Character names and annotations.....	43
27.1	Entity names	43
27.2	Name formation	43
27.3	Single name.....	44
27.4	Name immutability	44
27.5	Name uniqueness.....	44
27.6	Character names for CJK ideographs.....	45
27.7	Character names for Tangut ideographs	45
27.8	Character names for Nüshu characters	45
27.9	Character names for Khitan Small Script characters	46
27.10	Character names for Hangul syllables.....	46
28	Named UCS Sequence Identifiers	47
29	Structure of the Basic Multilingual Plane.....	49
30	Structure of the Supplementary Multilingual Plane for scripts and symbols (SMP)	51
31	Structure of the Supplementary Ideographic Plane (SIP).....	54
32	Structure of the Tertiary Ideographic Plane (TIP)	55
33	Structure of the Supplementary Special-purpose Plane (SSP).....	56
34	Code charts and lists of character names	57
34.1	General	57
34.2	Code chart.....	57
34.3	Character names list.....	57
34.4	Summary of standardized variation sequences	58
34.5	Code charts and lists of character names	58
Annex A (normative)	Collections of graphic characters for subsets	2743
A.1	Collections of coded graphic characters	2743
A.2	Blocks lists	2750
A.3	Fixed collections of the whole UCS (except Unicode collections)	2753
A.4	CJK collections.....	2756
A.5	Other collections	2757
A.6	Unicode collections	2761
Annex B (normative)	List of combining characters.....	2763
Annex C (normative)	Transformation format for planes 01 to 10 of the UCS (UTF-16)	2764
Annex D (normative)	UCS Transformation Format 8 (UTF-8).....	2765
Annex E (normative)	Mirrored characters in bidirectional context.....	2766
Annex F (informative)	Format characters	2767
F.1	General format characters	2767
F.2	Script-specific format characters	2769
F.3	Interlinear annotation characters	2770
F.4	Subtending format characters	2770
F.5	Shorthand format characters.....	2771
F.6	Invisible mathematical operators	2771

ISO/IEC 10646:2020 (E)

F.7	Western musical symbols	2771
F.8	Language tagging using Tag characters.....	2772
Annex G (informative)	Alphabetically sorted list of character names.....	2774
Annex H (informative)	The use of “signatures” to identify UCS	2775
Annex I (informative)	Ideographic description characters	2776
I.1	General	2776
I.2	Syntax of an ideographic description sequence.....	2776
I.3	Individual definitions of the ideographic description characters.....	2777
Annex J (informative)	Recommendation for combined receiving/originating devices with internal storage	2779
Annex K (informative)	Notations of octet value representations	2780
Annex L (informative)	Character naming guidelines	2781
Annex M (informative)	Sources of characters.....	2784
Annex N (informative)	External references to character repertoires	2785
N.1	Methods of reference to character repertoires and their coding.....	2785
N.2	Identification of ASN.1 character abstract syntaxes	2785
N.3	Identification of ASN.1 character transfer syntaxes.....	2786
Annex P (informative)	Additional information on CJK Unified ideographs	2787
Annex Q (informative)	Code mapping table for Hangul syllables	2790
Annex R (informative)	Names of Hangul syllables	2791
Annex S (informative)	Procedure for the unification and arrangement of CJK ideographs	2792
S.1	Unification procedure	2792
S.2	Arrangement procedure.....	2796
S.3	Source separation examples	2796
S.4	Non-unification examples.....	2801
Annex T (informative)	Language tagging using Tag Characters	2803
Annex U (informative)	Characters in identifiers	2804

ISO/IEC 10646:2020 (E)

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.

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 ISO documents 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) or the IEC list of patent declarations received (see <http://patents.iec.ch>).

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

For an explanation of 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 www.iso.org/iso/foreword.html.

This document was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 2, *Coded character sets*.

This sixth edition of ISO/IEC 10646 cancels and replaces the fifth edition (ISO/IEC 10646:2017), which has been technically revised. It also incorporates ISO/IEC 10646:2017/Amd 1:2019 and ISO/IEC 10646:2017/Amd 2:2019.

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

- New scripts covered: Chorasman, Dives Akuru, Dogra, Elymaic, Gunjala Gondi, Hanifi Rohingya, Khitan Small Script, Makasar, Medefaidrin, Nandinagari, Nyiakeng Puachue Hmong, Old Sogdian, Sogdian, Yezidi, Wancho;
- Existing scripts significantly extended: Georgian, CJK Unified Ideographs (Extension G);
- New symbol sets: Chess Symbols, Symbols for Legacy Computing;
- New set of Emoji symbols.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

ISO/IEC 10646:2020 (E)

Introduction

This document 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.

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 13.2.

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.

Information technology — Universal Coded Character Set (UCS)

1 Scope

This document

- specifies the architecture of the UCS;
- defines terms used for the UCS;
- describes the general structure of the UCS codespace;
- specifies the assigned planes of the UCS: the Basic Multilingual Plane (BMP) 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.

NOTE – The determination of suitability of these characters for use as identifiers in programming languages is not specified by this document but can be found in an external reference. See Annex U.

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 *Information technology — Character code structure and extension techniques*.

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

Unicode Standard Annex, UAX #9, *The Unicode Bidirectional Algorithm*:

<http://www.unicode.org/reports/tr9/tr9-42.html>

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

<http://www.unicode.org/reports/tr15/tr15-50.html>

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

<http://www.unicode.org/reports/tr37/tr37-12.html>

Unicode Standard Version 13.0, *Chapter 4, Character Properties*

<http://www.unicode.org/versions/Unicode13.0.0/ch04.pdf>

Section 4.3, *Combining Classes – Normative*

Section 4.5, *General Category – Normative*

Section 4.7, *Bidi Mirrored – Normative*

Unicode Standard Version 12.1, *Age Property*:

<https://www.unicode.org/Public/13.0.0/ucd/DerivedAge.txt>