

This is a preview - click here to buy the full publication

INTERNATIONAL STANDARD

ISO/IEC 10646

Fifth edition
2017-12

Information technology — Universal Coded Character Set (UCS)

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

Withdrawn



Reference number
ISO/IEC 10646:2017(E)

© ISO/IEC 2017

Withdrawn



COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2017, Published in Switzerland

All rights reserved. Unless otherwise specified, 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
Ch. de Blandonnet 8 • CP 401
CH-1214 Vernier, Geneva, Switzerland
Tel. +41 22 749 01 11
Fax +41 22 749 09 47
copyright@iso.org
www.iso.org

CONTENTS

| | |
|--|------|
| Foreword | vii |
| Introduction | viii |
| 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 General structure of the UCS..... | 9 |
| 6 Basic structure and nomenclature | 10 |
| 6.1 Structure..... | 10 |
| 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 |

ISO/IEC 10646:2017 (E)

| | | |
|------|---|----|
| 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 | 28 |
| 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 | 37 |
| 24 | Source references for Tangut ideographs..... | 37 |
| 24.1 | List of source references..... | 37 |
| 24.2 | Source reference file for Tangut ideographs | 38 |
| 24.3 | Source reference presentation for Tangut ideographs..... | 39 |

| | | |
|---------|--|------|
| 25 | Source references for Nüshu characters..... | 39 |
| 25.1 | List of source references..... | 39 |
| 25.2 | Source reference file for Nüshu characters | 39 |
| 26 | Character names and annotations..... | 40 |
| 26.1 | Entity names | 40 |
| 26.2 | Name formation | 40 |
| 26.3 | Single name..... | 41 |
| 26.4 | Name immutability | 41 |
| 26.5 | Name uniqueness..... | 41 |
| 26.6 | Character names for CJK ideographs..... | 42 |
| 26.7 | Character names for Tangut ideographs | 42 |
| 26.8 | Character names for Nüshu characters | 42 |
| 26.9 | Character names for Hangul syllables | 43 |
| 27 | Named UCS Sequence Identifiers | 44 |
| 28 | Structure of the Basic Multilingual Plane..... | 46 |
| 29 | Structure of the Supplementary Multilingual Plane for scripts and symbols (SMP)..... | 48 |
| 30 | Structure of the Supplementary Ideographic Plane (SIP)..... | 51 |
| 31 | Structure of the Tertiary Ideographic Plane (TIP) | 51 |
| 32 | Structure of the Supplementary Special-purpose Plane (SSP)..... | 51 |
| 33 | Code charts and lists of character names | 52 |
| 33.1 | General | 52 |
| 33.2 | Code chart..... | 52 |
| 33.3 | Character names list..... | 52 |
| 33.4 | Summary of standardized variation sequences | 53 |
| 33.5 | Code charts and lists of character names..... | 54 |
| Annex A | (normative) Collections of graphic characters for subsets | 2611 |
| A.1 | Collections of coded graphic characters..... | 2611 |
| A.2 | Blocks lists..... | 2617 |
| A.3 | Fixed collections of the whole UCS (except Unicode collections) | 2620 |
| A.4 | CJK collections..... | 2623 |
| A.5 | Other collections | 2624 |
| A.6 | Unicode collections | 2628 |
| Annex B | (normative) List of combining characters..... | 2629 |
| Annex C | (normative) Transformation format for planes 01 to 10 of the UCS (UTF-16) | 2630 |
| Annex D | (normative) UCS Transformation Format 8 (UTF-8)..... | 2631 |
| Annex E | (normative) Mirrored characters in bidirectional context..... | 2632 |
| Annex F | (informative) Format characters | 2633 |
| F.1 | General format characters | 2633 |
| F.2 | Script-specific format characters | 2635 |
| F.3 | Interlinear annotation characters | 2636 |
| F.4 | Subtending format characters | 2636 |
| F.5 | Shorthand format characters..... | 2637 |
| F.6 | Invisible mathematical operators | 2637 |

ISO/IEC 10646:2017 (E)

| | | |
|-----------------------|---|------|
| F.7 | Western musical symbols | 2637 |
| F.8 | Language tagging using Tag characters | 2638 |
| Annex G (informative) | Alphabetically sorted list of character names | 2640 |
| Annex H (informative) | The use of “signatures” to identify UCS | 2641 |
| Annex I (informative) | Ideographic description characters | 2642 |
| I.1 | General | 2642 |
| I.2 | Syntax of an ideographic description sequence | 2642 |
| I.3 | Individual definitions of the ideographic description characters | 2643 |
| Annex J (informative) | Recommendation for combined receiving/originating devices with internal storage | 2645 |
| Annex K (informative) | Notations of octet value representations | 2646 |
| Annex L (informative) | Character naming guidelines | 2647 |
| Annex M (informative) | Sources of characters | 2650 |
| Annex N (informative) | External references to character repertoires | 2674 |
| N.1 | Methods of reference to character repertoires and their coding | 2674 |
| N.2 | Identification of ASN.1 character abstract syntaxes | 2674 |
| N.3 | Identification of ASN.1 character transfer syntaxes | 2675 |
| Annex P (informative) | Additional information on CJK Unified Ideographs | 2676 |
| Annex Q (informative) | Code mapping table for Hangul syllables | 2679 |
| Annex R (informative) | Names of Hangul syllables | 2680 |
| Annex S (informative) | Procedure for the unification and arrangement of CJK ideographs | 2681 |
| S.1 | Unification procedure | 2681 |
| S.2 | Arrangement procedure | 2685 |
| S.3 | Source separation examples | 2685 |
| S.4 | Non-unification examples | 2690 |
| Annex T (informative) | Language tagging using Tag Characters | 2692 |
| Annex U (informative) | Characters in identifiers | 2693 |

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.

ISO/IEC 10646:2017 (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 130 000 characters from the world's scripts.

Withdrawn

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>

ISO/IEC 10646:2017 (E)

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>

Withdrawn