INTERNATIONAL STANDARD

ISO/IEC 10646

Second edition
2011-03-15

Information technology — Universal Coded Character Set (UCS)

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

Withdrawn

Reference number
ISO/IEC 10646:2011(E)

© ISO/IEC 2011

PDF disclaimer

This PDF file may contain embedded typefaces. In accordance with Adobe's licensing policy, this file may be printed or viewed but shall not be edited unless the typefaces which are embedded are licensed to and installed on the computer performing the editing. In downloading this file, parties accept therein the responsibility of not infringing Adobe's licensing policy. The ISO Central Secretariat accepts no liability in this area.

Adobe is a trademark of Adobe Systems Incorporated.

Details of the software products used to create this PDF file can be found in the General Info relative to the file; the PDF-creation parameters were optimized for printing. Every care has been taken to ensure that the file is suitable for use by ISO member bodies. In the unlikely event that a problem relating to it is found, please inform the Central Secretariat at the address given below.

withdrawn

COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2011

All rights reserved. Unless otherwise specified, no part of this publication may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying and microfilm, without permission in writing from either ISO at the address below or ISO's member body in the country of the requester.

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

 © ISO/IEC 2011 – All rights reserved

CONTENTS

Foreword 7
Introduction 8
1 Scope 9
2 Conformance 9
2.1 General 9
2.2 Conformance of information interchange 9
2.3 Conformance of devices 10
3 Normative references 10
4 Terms and definitions 11
5 General structure of the UCS 17
6 Basic structure and nomenclature 17
6.1 Structure 17
6.2 Coding of characters 19
6.3 Types of code points 19
6.4 Naming of characters 20
6.5 Short identifiers for code points (UIDs) 20
6.6 UCS Sequence Identifiers 21
6.7 Octet sequence identifiers 21
7 Revision and updating of the UCS 22
8 Subsets 22
8.1 Limited subset 22
8.2 Selected subset 22
9 UCS encoding forms 22
9.1 UTF-8 22
9.2 UTF-16 23
9.3 UTF-32 (UCS-4) 24
10 UCS Encoding schemes 24
10.1 UTF-8 24
10.2 UTF-16BE 24
10.3 UTF-16LE 24
10.4 UTF-16 24
10.5 UTF-32BE 25
10.6 UTF-32LE 25
10.7 UTF-32 25
11 Use of control functions with the UCS 25
12 Declaration of identification of features 26
12.1 Purpose and context of identification 26
12.2 Identification of a UCS encoding form 27
12.3 Identification of subsets of graphic characters 27
12.4 Identification of control function set 27
12.5 Identification of the coding system of ISO/IEC 2022 28
13 Structure of the code charts and lists 28

Withdrawn

© ISO/IEC 2011 – All rights reserved

14 Block and collection names 29
14.1 Block names 29
14.2 Collection names 29
15 Mirrored characters in bidirectional context 29
15.1 Mirrored characters 29
15.2 Directionality of bidirectional text 29
16 Special characters 30
16.1 Space characters 30
16.2 Currency symbols 30
16.3 Format Characters 30
16.4 Ideographic description characters 31
16.5 Variation selectors and variation sequences 31
17 Presentation forms of characters 34
18 Compatibility characters 34
19 Order of characters 34
20 Combining characters 35
20.1 Order of combining characters 35
20.2 Combining class and canonical ordering 35
20.3 Appearance in code charts 35
20.4 Alternate coded representations 35
20.5 Multiple combining characters 35
20.6 Collections containing combining characters 36
20.7 Combining Grapheme Joiner 36
21 Normalization forms 36
22 Special features of individual scripts and symbol repertoires 37
22.1 Hangul syllable composition method 37
22.2 Features of scripts used in India and some other South Asian countries 37
22.3 Byzantine musical symbols 38
22.4 Source references for pictographic symbols 38
23 Source references for CJK Ideographs 38
23.1 List of source references 38
23.2 Source references for CJK Unified Ideographs 41
23.3 Source reference presentation for CJK Unified Ideographs 42
23.4 Source references for CJK Compatibility Ideographs 43
23.5 Source references presentation for CJK Compatibility Ideographs 44
24 Character names and annotations 44
24.1 Entity names 44
24.2 Name formation 44
24.3 Single name 45
24.4 Name immutability 45
24.5 Name uniqueness 45
24.6 Character names for CJK Ideographs 46
24.7 Character names for Hangul syllables 46
25 Named UCS Sequence Identifiers 48

Withdrawn

 © ISO/IEC 2011 – All rights reserved

26 Structure of the Basic Multilingual Plane 50
27 Structure of the Supplementary Multilingual Plane for scripts and symbols (SMP) 52
28 Structure of the Supplementary Ideographic Plane (SIP) 54
29 Structure of the Tertiary Ideographic Plane (TIP) 54
30 Structure of the Supplementary Special-purpose Plane (SSP) 54
31 Code charts and lists of character names 55
31.1 Code chart 55
31.2 Character names list 55
31.3 Pointers to code charts and lists of character names 56
Annex A (normative) Collections of graphic characters for subsets 1599
A.1 Collections of coded graphic characters 1599
A.2 Blocks lists 1604
A.3 Fixed collections of the whole UCS (except Unicode collections) 1606
A.4 CJK collections 1609
A.5 Other collections 1610
A.6 Unicode collections 1613
Annex B (normative) List of combining characters 1622
Annex C (normative) Transformation format for planes 01 to 10 of the UCS (UTF-16) 1623
Annex D (normative) UCS Transformation Format 8 (UTF-8) 1624
Annex E (normative) Mirrored characters in bidirectional context 1625
Annex F (informative) Format characters 1626
F.1 General format characters 1626
F.2 Script-specific format characters 1628
F.3 Interlinear annotation characters 1629
F.4 Subtending format characters 1629
F.5 Contiguity operators 1629
F.6 Western musical symbols 1630
F.7 Language tagging using Tag characters 1631
Annex G (informative) Alphabetically sorted list of character names 1633
Annex H (informative) The use of –signatures" to identify UCS 1634
Annex I (informative) Ideographic description characters 1635
Annex J (informative) Recommendation for combined receiving/originating devices with internal storage 1638
Annex K (informative) Notations of octet value representations 1639
Annex L (informative) Character naming guidelines 1640
Annex M (informative) Sources of characters 1643
Annex N (informative) External references to character repertoires 1660
N.1 Methods of reference to character repertoires and their coding 1660
N.2 Identification of ASN.1 character abstract syntaxes 1660
N.3 Identification of ASN.1 character transfer syntaxes 1661
Annex P (informative) Additional information on CJK Unified Ideographs 1662
Annex Q (informative) Code mapping table for Hangul syllables 1663
Annex R (informative) Names of Hangul syllables 1664

Withdrawn

© ISO/IEC 2011 – All rights reserved

Annex S (informative) Procedure for the unification and arrangement of CJK Ideographs 1665
S.1 Unification procedure 1665
S.2 Arrangement procedure 1669
S.3 Source code separation examples 1669
S.4 Non-unification examples 1674
Annex T (informative) Language tagging using Tag Characters 1676
Annex U (informative) Characters in identifiers 1677

Withdrawn

 © ISO/IEC 2011 – All rights reserved

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.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of the joint technical committee is to prepare International Standards. Draft International Standards adopted by the joint technical committee are circulated to national bodies for voting. Publication as an International Standard requires approval by at least 75 % of the national bodies casting a vote.

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.

ISO/IEC 10646 was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 2, *Coded character sets*.

This second edition cancels and replaces the first edition (ISO/IEC 10646:2003), which has been technically revised. It also incorporates the Amendments ISO/IEC 10646:2003/Amd.1:2005, ISO/IEC 10646:2003/Amd.2:2006, ISO/IEC 10646:2003/Amd.3:2008, ISO/IEC 10646:2003/Amd.4:2008, ISO/IEC 10646:2003/Amd.5:2008, ISO/IEC 10646:2003/Amd.6:2009, and ISO/IEC 10646:2003/Amd.7:2010.

Withdrawn

© ISO/IEC 2011 – All rights reserved

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.

Withdrawn

 © ISO/IEC 2011 – All rights reserved

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 worldwide 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 6.0 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

Withdrawn

© ISO/IEC 2011 – All rights reserved

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/tr9-23.html

Unicode Standard Annex, UAX #15, *Unicode Normalization Forms:*
http://www.unicode.org/reports/tr15/tr15-33.html

Unicode Technical Standard, *UTS #37, Unicode Ideographic Variation Database:*
http://www.unicode.org/reports/tr37/tr37-5.html

 © ISO/IEC 2011 – All rights reserved

Unicode Standard Version 6.0, *Chapter 4, Character Properties*
http://www.unicode.org/versions/Unicode6.0.0/
Section 4.3, Combining Classes – Normative
Section 4.5, General Category – Normative
Section 4.7, Bidi Mirrored – Normative

Withdrawn

© ISO/IEC 2011 – All rights reserved