
**Information technology — Universal
Multiple-Octet Coded Character Set (UCS)**

*Technologies de l'information — Jeu universel de caractères codés sur
plusieurs octets (JUC)*

Withdrawal

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Withdrawn

© ISO/IEC 2003

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

Contents	Page
1 Scope	1
2 Conformance.....	1
3 Normative references.....	2
4 Terms and definitions.....	2
5 General structure of the UCS.....	4
6 Basic structure and nomenclature.....	5
7 General requirements for the UCS.....	9
8 The Basic Multilingual Plane	9
9 Supplementary planes	10
10 Private use groups, planes, and zones	10
11 Revision and updating of the UCS	10
12 Subsets	10
13 Coded representation forms of the UCS	11
14 Implementation levels	11
15 Use of control functions with the UCS.....	11
16 Declaration of identification of features	12
17 Structure of the code tables and lists	13
18 Block names.....	13
19 Characters in bi-directional context.....	14
20 Special characters.....	14
21 Presentation forms of characters	17
22 Compatibility characters.....	18
23 Order of characters	18
24 Normalization forms	18
25 Combining characters	18
26 Special features of individual scripts.....	20
27 Source references for CJK Ideographs.....	20
28 Character names and annotations	23
29 Structure of the Basic Multilingual Plane.....	25
30 Structure of the Supplementary Multilingual Plane for Scripts and symbols.....	27
31 Structure of the Supplementary Ideographic Plane	28
32 Supplementary Special-purpose Plane.....	28
33 Code tables and lists of character names	28

NOTE The code tables and lists of character names are given on pages 29-1348. They are contained in separate files which are accessed by clicking on the appropriate highlighted text in Clause 33.

Annexes

A (normative) Collections of graphic characters for subsets	1349
B (normative) List of combining characters	1358
C (normative) Transformation format for 16 planes of Group 00 (UTF-16)	1364

D (normative) UCS Transformation Format 8 (UTF-8).....	1367
E (informative) Mirrored characters in Arabic bi-directional context.....	1371
F (informative) Alternate format characters.....	1374
G (informative) Alphabetically sorted list of character names.....	1379
H (informative) The use of “signatures” to identify UCS.....	1380
J (informative) Recommendation for combined receiving/originating devices with internal storage	1381
K (informative) Notations of octet value representations	1382
L (informative) Character naming guidelines	1383
M (informative) Sources of characters	1386
N (informative) External references to character repertoires	1390
P (informative) Additional information on characters	1392
Q (informative) Code mapping table for Hangul syllables.....	1395
R (informative) Names of Hangul syllables.....	1396
S (informative) Procedure for the unification and arrangement of CJK Ideographs.....	1408
T (informative) Language tagging using Tag Characters.....	1416
U (informative) Usage of musical symbols.....	1418

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 first edition of ISO/IEC 10646 cancels and replaces ISO/IEC 10646-1:2000 and ISO/IEC 10646-2:2001. It also incorporates ISO/IEC 10646-1:2000/Amd.1:2002.

Introduction

ISO/IEC 10646 specifies the Universal Multiple-Octet 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. ISO/IEC 10646 has been widely adopted in new Internet protocols and implemented in modern operating systems and computer languages. This edition covers over 95 000 characters from the world's scripts.

ISO/IEC 10646 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:

- CJKU_SR.txt
- CJKC_SR.txt
- Allnames.txt
- HangulX.txt
- HangulSy.txt

Information technology — Universal Multiple-Octet Coded Character Set (UCS)

1 Scope

ISO/IEC 10646 specifies the Universal Multiple-Octet 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 document:

- specifies the architecture of ISO/IEC 10646,
- defines terms used in ISO/IEC 10646,
- describes the general structure of the coded character set;
- 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) 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 of the BMP, SMP, SIP, SSP and their coded representations;
- specifies the four-octet (32-bit) canonical form of the UCS: UCS-4;
- specifies a two-octet (16-bit) BMP form of the UCS: UCS-2;
- specifies the coded representations for control functions;
- specifies the management of future additions to this coded character set.

The UCS is a coding system different from that specified in ISO/IEC 2022. The method to designate UCS from ISO/IEC 2022 is specified in clause 16.2.

A graphic character will be assigned only one code position in the standard, located either in the BMP or in one of the supplementary planes.

NOTE – The Unicode Standard, Version 4.0 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 ISO/IEC 10646, the characters themselves shall not be covered by these conformance requirements.

2.2 Conformance of information interchange

A coded-character-data-element (CC-data-element) within coded information for interchange is in conformance with ISO/IEC 10646 if

- a) all the coded representations of graphic characters within that CC-data-element conform to clauses 6 and 7, to an identified form chosen from clause 13 or annex C or annex D, and to an identified implementation level chosen from clause 14;
- b) all the graphic characters represented within that CC-data-element are taken from those within an identified subset (see clause 12);
- c) all the coded representations of control functions within that CC-data-element conform to clause 15.

A claim of conformance shall identify the adopted form, the adopted implementation level 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 ISO/IEC 10646 if it conforms to the requirements of item a) below, and either or both of items b) and c).

NOTE – The term device is defined (in 4.18) as a component of information processing equipment which can transmit and/or receive coded information within CC-data-elements. A device may be a conventional input/output device, or a process such as an application program or gateway function.

A claim of conformance shall identify the document that contains the description specified in a) below, and shall identify the adopted form(s), the adopted implementation level, the adopted subset (by means of a list of collections and/or characters), and the selection of control functions adopted in accordance with clause 15.

- a) **Device description:** A device that conforms to ISO/IEC 10646 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 sub-clauses 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 CC-data-element in accordance with the adopted form and implementation level.
- c) **Receiving device:** A receiving device shall be capable of receiving and interpreting any coded representation of characters that are within a CC-data-element in accordance with the adopted form and implementation level, and shall make any corresponding characters from the adopted subset available to the user in such a way that the user can identify them.

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 – An indication to the user may consist of making available the same character to represent all characters not in the adopted subset, or providing a distinctive audible or visible signal when appropriate to the type of user.

NOTE 2 – See also annex J for receiving devices with re-transmission 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, Version 4.0.0, 2003-04-17.

Unicode Standard Annex, UAX#15, Unicode Normalization Forms, Version 4.0.0, 2003-04-17.