
Information technology — Universal Coded Character Set (UCS)

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



COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2014

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

Foreword.....	vii
Introduction	viii
1 Scope	1
2 Conformance	1
2.1 General	1
2.2 Conformance of information interchange.....	1
2.3 Conformance of devices	2
3 Normative references	2
4 Terms and definitions	3
5 General structure of the UCS	9
6 Basic structure and nomenclature	9
6.1 Structure.....	9
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
12.3 Identification of subsets of graphic characters.....	19

ISO/IEC 10646:2014 (E)

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	29
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	36
24	Character names and annotations	37
24.1	Entity names	37
24.2	Name formation.....	37
24.3	Single name	38
24.4	Name immutability.....	38
24.5	Name uniqueness	38

24.6	Character names for CJK Ideographs	39
24.7	Character names for Hangul syllables	39
25	Named UCS Sequence Identifiers	41
26	Structure of the Basic Multilingual Plane	42
27	Structure of the Supplementary Multilingual Plane for scripts and symbols (SMP)	44
28	Structure of the Supplementary Ideographic Plane (SIP)	46
29	Structure of the Tertiary Ideographic Plane (TIP)	46
30	Structure of the Supplementary Special-purpose Plane (SSP)	46
31	Code charts and lists of character names	47
31.1	General	47
31.2	Code chart	47
31.3	Character names list	47
31.4	Summary of standardized variation sequences	48
31.5	Pointers to code charts and lists of character names	48
Annex A	(normative) Collections of graphic characters for subsets	2381
A.1	Collections of coded graphic characters	2381
A.2	Blocks lists	2387
A.3	Fixed collections of the whole UCS (except Unicode collections)	2389
A.4	CJK collections	2393
A.5	Other collections	2393
A.6	Unicode collections	2397
Annex B	(normative) List of combining characters	2410
Annex C	(normative) Transformation format for planes 01 to 10 of the UCS (UTF-16)	2411
Annex D	(normative) UCS Transformation Format 8 (UTF-8)	2412
Annex E	(normative) Mirrored characters in bidirectional context	2413
Annex F	(informative) Format characters	2414
F.1	General format characters	2414
F.2	Script-specific format characters	2416
F.3	Interlinear annotation characters	2417
F.4	Subtending format characters	2417
F.5	Shorthand format characters	2418
F.6	Invisible mathematical operators	2418
F.7	Western musical symbols	2418
F.8	Language tagging using Tag characters	2419
Annex G	(informative) Alphabetically sorted list of character names	2421
Annex H	(informative) The use of “signatures” to identify UCS	2422
Annex I	(informative) Ideographic description characters	2423
I.1	General	2423
I.2	Syntax of an ideographic description sequence	2423
I.3	Individual definitions of the ideographic description characters	2423
Annex J	(informative) Recommendation for combined receiving/originating devices with internal storage	2426
Annex K	(informative) Notations of octet value representations	2427

ISO/IEC 10646:2014 (E)

Annex L (informative) Character naming guidelines	2428
Annex M (informative) Sources of characters	2431
Annex N (informative) External references to character repertoires	2449
N.1 Methods of reference to character repertoires and their coding	2449
N.2 Identification of ASN.1 character abstract syntaxes	2449
N.3 Identification of ASN.1 character transfer syntaxes	2450
Annex P (informative) Additional information on CJK Unified Ideographs	2451
Annex Q (informative) Code mapping table for Hangul syllables	2454
Annex R (informative) Names of Hangul syllables	2455
Annex S (informative) Procedure for the unification and arrangement of CJK Ideographs	2456
S.1 Unification procedure	2456
S.2 Arrangement procedure	2459
S.3 Source separation examples	2460
S.4 Non-unification examples	2465
Annex T (informative) Language tagging using Tag Characters	2466
Annex U (informative) Characters in identifiers	2467

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 meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the WTO principles in the Technical Barriers to Trade (TBT) see the following URL: [Foreword - Supplementary information](#)

The committee responsible for this document is ISO/IEC JTC 1, *Information technology*, SC 2, *Coded character sets*.

This fourth edition cancels and replaces the third edition (ISO/IEC 10646:2012), which has been technically revised. It also incorporates the Amendment ISO/IEC 10646:2012/Amd.1:2013.

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 120 000 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
- UCSVariants.txt
- CJKSrc.txt
- NUSI.txt
- JIExt.txt
- Allnames.txt
- HangulSy.txt.

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 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

- 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 Clause 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 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 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/tr9-31.html>.

Unicode Standard Annex, UAX #15, *Unicode Normalization Forms*:
<http://www.unicode.org/reports/tr15/tr15-41.html>.

Unicode Technical Standard, UTS #37, *Ideographic Variation Database*:
<http://www.unicode.org/reports/tr37/tr37-8.html>.

Unicode Standard Version 6.2, *Chapter 4, Character Properties*
<http://www.unicode.org/versions/Unicode7.0.0/ch04.pdf>
Section 4.3, Combining Classes – Normative

Section 4.5, General Category – Normative

Section 4.7, Bidi Mirrored – Normative

4 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

4.1

base character

graphic character which is not a combining character

NOTE 1 – 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 – A base character typically does not graphically combine with preceding characters. There are exceptions for some complex writing systems.

4.2

Basic Multilingual Plane

BMP

plane 00 of the UCS codespace

4.3

block

contiguous range of code points to which a set of characters that share common characteristics, such as a script, are allocated; a block does not overlap another block; one or more of the code points within a block may have no character allocated to them

4.4

canonical form

form with which characters of this coded character set are specified using a single code point within the UCS codespace

NOTE – The canonical form is not to be confused with an encoding form which describes the relationship between UCS code points and one or several code units (see 4.23).

4.5

character

member of a set of elements used for the organization, control, or representation of textual data

NOTE – A graphic symbol can be represented by a sequence of one or several coded characters.

4.6

character boundary

(code unit sequence) demarcation between the last code unit of a coded character and the first code unit of the next coded character

4.7

code chart

code table

rectangular array showing the representation of coded characters allocated within a range of the UCS codespace

4.8

coded character

association between a character and a code point

4.9

coded character set

set of coded characters

4.10

code point

code position

value in the UCS codespace

4.11

code unit

minimal bit combination that can represent a unit of encoded text for processing or interchange

NOTE – Examples of code units are octets (8-bit code units) used in the UTF-8 encoding form, 16-bit code units in the UTF-16 encoding form, and 32-bit code units in the UTF-32 encoding form.

4.12

code unit sequence

CC-data-element

coded-character-data-element

element of interchanged information that is specified to consist of a sequence of code units, in accordance with one or more identified standards for coded character sets

NOTE 1 – Such sequence can contain code units associated with any type of code points.

NOTE 2 – Since its second edition: ISO/IEC 10646:2011, this International Standard does not use implementation levels. Its definition of code unit sequence corresponds to the former unrestricted implementation level 3. Other definitions of code unit sequence, previously known as level 1 and 2, are deprecated. To maintain compatibility with these previous editions, in the context of identification of coded representation in International Standards such as ISO/IEC 8824 and ISO/IEC 8825, the concept of implementation level can still be referenced as 'Implementation level 3'. See Annex N.

4.13

collection

numbered and named set of entities

NOTE 1 – For a non extended collection, these entities consist only of those coded characters whose code points lie within one or more identified ranges (see also 4.25 for extended collection).

NOTE 2 – If any of the identified ranges include code points to which no character is allocated, the repertoire of the collection will change if an additional character is assigned to any of those code points at a future amendment of this International Standard. However it is intended that the collection number and name will remain unchanged in future editions of this International Standard.

4.14

combining character

character which has General Category values of Spacing Combining Mark (Mc), Non Spacing Mark (Mn), and Enclosing Mark (Me)

NOTE – These characters are intended for combination with the preceding non-combining graphic character, or with a sequence of combining characters preceded by a non-combining character (see also 4.17).

4.15

combining class

value associated with each combining character determining its typographical interaction and its canonical ordering within a sequence of combining characters

4.16

compatibility character

graphic character included as a coded character of this International Standard primarily for compatibility with existing coded character sets

4.17

composite sequence

sequence of graphic characters consisting of a base character followed by one or more combining characters, ZERO WIDTH JOINER, or ZERO WIDTH NON-JOINER (see also 4.14)

NOTE 1 – A graphic symbol for a composite sequence generally consists of the combination of the graphic symbols of each character in the sequence.

NOTE 2 – A composite sequence can be used to represent characters not encoded in the repertoire of this International Standard.

4.18

control character

control function the coded representation of which consists of a single code point

NOTE – Although control characters are often 'named' using terms such as DELETE, FORM FEED, ESC, these qualifiers do not correspond to formal character names. See Clause 11 for a list of the long names used by ISO/IEC 6429 in association with the control characters.

4.19

control function

action that affects the recording, processing, transmission, or interpretation of data, and that is represented by a code unit sequence

4.20

decomposition mapping

mapping from a character to a sequence of one or more characters that is a canonical or compatibility equivalent

4.21

default state

state that is assumed when no state has been explicitly specified

NOTE – See F.2.1, F.2.2, and F.2.3.

4.22

device

component of information processing equipment which can transmit and/or receive coded information within code unit sequences

NOTE – It may be an input/output device in the conventional sense, or a process such as an application program or gateway function.

4.23

encoding form

form that determines how each UCS code point for a UCS character is to be expressed as one or more code units used by the encoding form

NOTE – This International Standard specifies UTF-8, UTF-16, and UTF-32.

4.24

encoding scheme

scheme that specifies the serialization of the code units from the encoding form into octets

NOTE – Some of the UCS encoding schemes have the same labels as the UCS encoding form. However they are used in different contexts. UCS encoding forms refer to in-memory and application interface representation of textual data. UCS encoding schemes refer to octet-serialized textual data.

4.25

extended collection

collection for which the entities can also consist of sequences of code points that are in Normalization Form C (NFC)

NOTE 1 – Some collections such as 3 LATIN EXTENDED-A, 4 LATIN EXTENDED-B, 15 ARABIC EXTENDED, and many more, have the term 'extended' in their name. This does not make them extended collections.

NOTE 2 – See Clause 21 for discussion of Normalization Form C.

NOTE 3 – The sequences of code points are referenced by Named UCS Sequence Identifiers (NUSI) (see Clause 25).

4.26

fixed collection

collection in which every code point within the identified range(s) has a character allocated to it, and which is intended to remain unchanged in future editions of this International Standard

4.27

format character

character whose primary function is to affect the layout or processing of characters around it

NOTE – A format character generally does not have a visible representation of its own.

4.28

General Category

GC

value assigned to each UCS code point which determines its major class, such as letter, punctuation, and symbol

NOTE 1 – Each value is defined as General Category property using a two-letter abbreviation in the Unicode Standard (see reference to the current Unicode Standard General Category in 3).

NOTE 2 – When referred as a group containing all GC values sharing the same first letter, the group may be described using the first letter only. For example, 'L' stands for all letters 'Lu', 'Ll', 'Lt', 'Lm', and 'Lo'.

4.29

graphic character

character, other than a control function or a format character, that has a visual representation normally handwritten, printed, or displayed

4.30

graphic symbol

visual representation of a graphic character or of a composite sequence

4.31

high-surrogate code point

code point in the range D800 to DBFF reserved for the use of UTF-16

4.32

high-surrogate code unit

16-bit code unit in the range D800 to DBFF used in UTF-16 as the leading code unit of a surrogate pair

NOTE – See 9.3.

4.33

ill-formed code unit sequence

UCS code unit sequence that purports to be in a UCS encoding form which does not conform to the specification of that encoding form

EXAMPLE – An unpaired surrogate code unit is an ill-formed code unit sequence.

4.34

ill-formed code unit sequence subset

non-empty subset of a code unit sequence X which does not contain any code unit which also belong to any minimal well-formed code unit sequence subset of X

NOTE – An ill-formed code unit sequence subset cannot overlap with a minimal well-formed code unit sequence.

4.35

interchange

transfer of character coded data from one user to another, using telecommunication means or interchangeable media

NOTE – Interchange implies data serialization and the use of a UCS encoding scheme.

4.36

interworking

process of permitting two or more systems, each employing different coded character sets, to meaningfully interchange character coded data

NOTE – Conversion between the two codes might be involved.

4.37**ISO/IEC 10646-1**

former subdivision of ISO/IEC 10646 containing the specification of the overall architecture and the Basic Multilingual Plane (BMP)

NOTE 1 – It is also referred to as Part 1 of ISO/IEC 10646.

NOTE 2 – There are a first and a second Edition of ISO/IEC 10646-1.

4.38**ISO/IEC 10646-2**

former subdivision of ISO/IEC 10646 containing the specification of the Supplementary Multilingual Plane (SMP), the Supplementary Ideographic Plane (SIP) and the Supplementary Special-purpose Plane (SSP)

NOTE 1 – It is also referred to as Part 2 of ISO/IEC 10646.

NOTE 2 – There is only a first edition of ISO/IEC 10646-2.

4.39**low-surrogate code point**

code point in the range DC00 to DFFF reserved for the use of UTF-16

4.40**low-surrogate code unit**

16-bit code unit in the range DC00 to DFFF used in UTF-16 as the trailing code unit of a surrogate pair

NOTE – See 9.3.

4.41**minimal well-formed code unit sequence**

well-formed code unit sequence that maps to a single UCS scalar value

4.42**mirrored character**

character whose image is mirrored horizontally in text that is laid out from right to left

4.43**octet**

8-bit code unit

NOTE – The value is expressed in hexadecimal notation from 00 to FF in this International Standard (see Annex K).

4.44**plane**

subdivision of the UCS codespace consisting of contiguous 65 536 code points beginning at a multiple of 65 536 which can be identified by a number from 00 to 10

NOTE – The UCS codespace contain 17 planes.

4.45**presentation**

process of writing, printing, or displaying a graphic symbol

4.46**presentation form**

(in the presentation of some scripts) form of a graphic symbol representing a character that depends on the position of the character relative to other characters

4.47**private use plane**

plane within this coded character set, the content of which is not specified in this International Standard

NOTE – Planes 0F and 10 are private use planes.

4.48

repertoire

specified set of characters that are represented in a coded character set

4.49

row

subdivision of a plane consisting of contiguous 256 code points beginning at a multiple of 256 which can be identified by a number from 00 to FF

4.50

script

set of graphic characters used for the written form of one or more languages

4.51

supplementary plane

plane other than Plane 00 of the UCS codespace

NOTE – A supplementary plane accommodates characters which have not been allocated to the Basic Multilingual Plane.

4.52

Supplementary Multilingual Plane for scripts and symbols

SMP

plane 01 of the UCS codespace

4.53

Supplementary Ideographic Plane

SIP

plane 02 of the UCS codespace

4.54

Supplementary Special-purpose Plane

SSP

plane 0E of the UCS codespace

4.55

surrogate pair

representation for a single character that consists of a sequence of two 16-bit code units, where the first value of the pair is a high-surrogate code unit and the second value is a low-surrogate code unit

4.56

Tertiary Ideographic Plane

TIP

plane 03 of the UCS codespace

4.57

UCS codespace

codespace consisting of the integers from 0 to 10 FFFF (hexadecimal) available for assigning the repertoire of the UCS characters

4.58

UCS scalar value

any UCS code point except high-surrogate and low-surrogate code points

4.59

unpaired surrogate code unit

code unit in a code unit sequence that is either a high-surrogate code unit that is not immediately followed by a low-surrogate unit, or a low-surrogate code unit that is not immediately preceded by a high-surrogate code unit

4.60**user**

person or other entity that invokes the service provided by a device

NOTE – This entity can be a process such as an application program if the “device” is a code converter or a gateway function.

4.61**well-formed code unit sequence**

UCS code unit sequence that purports to be in a UCS encoding form which conforms to the specification of that encoding form and contains no ill-formed code unit sequence subset

5 General structure of the UCS

The general structure of the Universal Coded Character Set (referred to hereafter as “this coded character set”) is described in this explanatory clause, and is illustrated in figure 1. The normative specification of the structure is given in the following clauses.

The canonical form of this coded character set – the way in which it is to be conceived – uses the UCS codespace which consists of the integers from 0 to 10FFFF.

This International Standard defines coded characters for the following planes:

- The Basic Multilingual Plane (BMP, Plane 00).
- The Supplementary Multilingual Plane for scripts and symbols (SMP, Plane 01).
- The Supplementary Ideographic Plane (SIP, Plane 02).
- The Supplementary Special-purpose Plane (SSP, Plane 0E).

The Tertiary Ideographic Plane (TIP, Plane 03) is reserved for ideographic characters and is currently empty. The planes from 04 to 0D are reserved for future standardization.

The planes 0F and 10 are reserved for private use.

Subsets of the coding space may be used in order to give a sub-repertoire of graphic characters.

6 Basic structure and nomenclature**6.1 Structure**

The Universal Coded Character Set as specified in this International Standard shall be regarded as a single entity made of 17 planes.

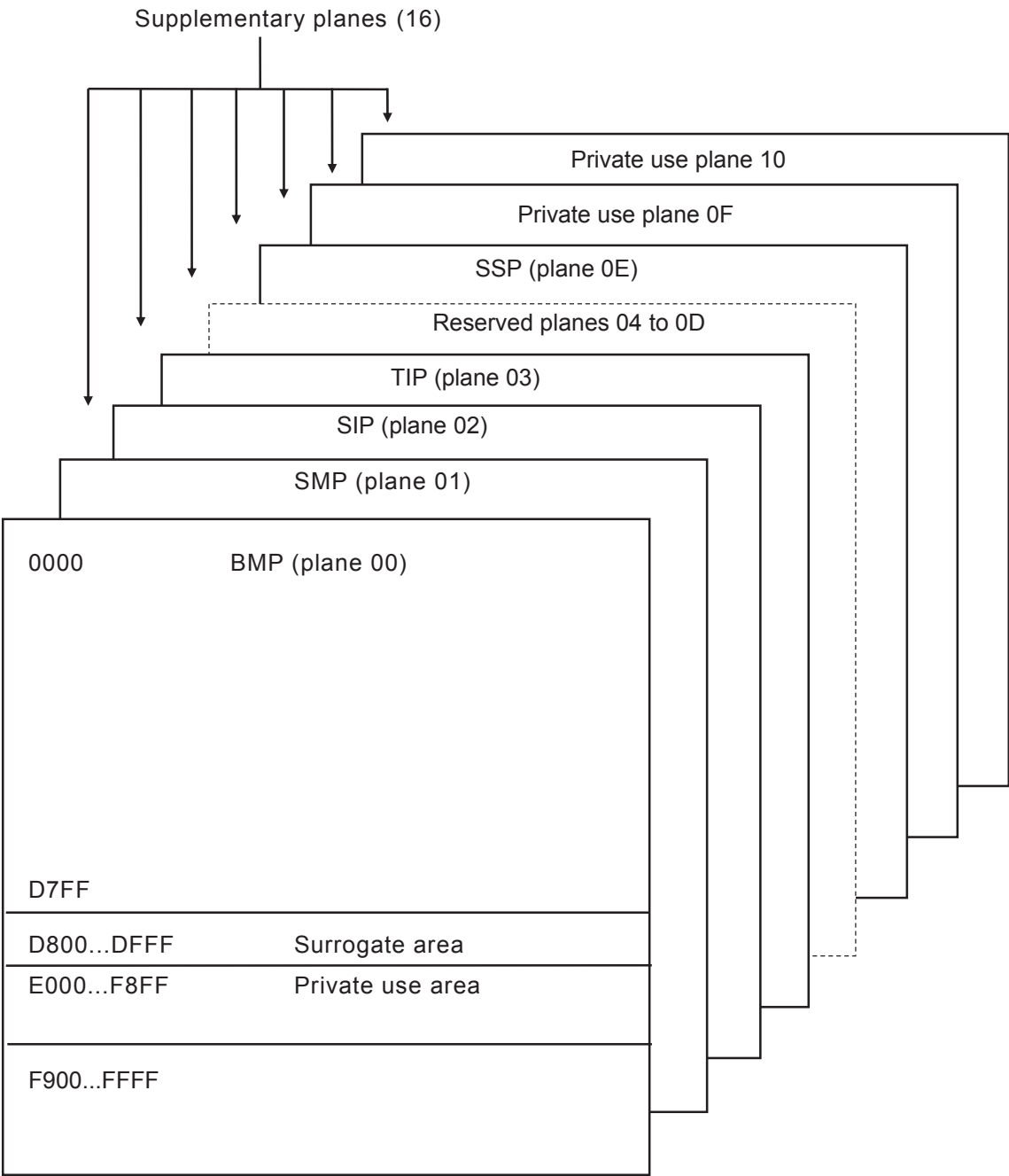


Figure 1 - Planes of the Universal Coded Character Set

6.2 Coding of characters

Each encoded character within the UCS codespace is represented by an integer between 0 and 10FFFF identified as a code point.

When a single character is to be identified in terms of its code point, it is represented by a six digit form of the integer such as

000030 for DIGIT ZERO
 000041 for LATIN CAPITAL LETTER A
 010000 for LINEAR B SYLLABLE B008 A

When referring to characters within plane 00, the leading two digits may be omitted; for characters within planes 01 to 0F, the leading digit may be omitted, such as

0030 for DIGIT ZERO
 0041 for LATIN CAPITAL LETTER A
 10000 for LINEAR B SYLLABLE B008 A

6.3 Types of code points

6.3.1 Classification

UCS code points are categorized in basic types, according to their General Category value. These values shall be determined according to the Unicode Standard General Category property (see Clause 3). The Table 1 summarizes the types:

Table 1: Type of code points

Basic Type	Brief Description	General Category	Character status	Code point status
Graphic	Letter, mark, number, punctuation, symbols, and spaces	L, M, N, P, S, Zs	Assigned to character	Assigned code point
Format	Invisible, but affects neighbouring characters	Cf, Zl, Zp		
Control	Control functions consisting of a single code point	Cc		
Private use	Usage defined by private agreement outside this standard	Co		
Surrogate	Permanently reserved for UTF-16	Cs	Not assigned to character	
Noncharacter	Permanently reserved for internal usage	Cn		
Reserved	Reserved for future assignment			
				Unassigned code point

Surrogate, noncharacter, and reserved code points are not assigned to characters and are subject to restriction in interchange. For example, surrogate code points do not have well-formed representations in any UCS encoding form.

6.3.2 Graphic characters

The same graphic character shall not be allocated to more than one code point. There are graphic characters with similar shapes in the coded character set; they are used for different purposes and have different character names.

6.3.3 Format characters

Format characters form a class of characters which are invisible but affect neighbouring characters.

6.3.4 Control characters

Code points 0000 to 001F, 007F to 009F in the BMP are reserved for control characters (see Clause 11).

6.3.5 Private use characters

Code points from E000 to F8FF in the BMP are reserved for private use. All code points of Plane 0F and Plane 10, except for FFFFE, FFFFF, 10FFFE, and 10FFFF are reserved for private use.

Private use characters are not constrained in any way by this International Standard. Private use characters can be used to provide user-defined characters. For example, this is a common requirement for users of ideographic scripts.

NOTE – For meaningful interchange of private use characters, an agreement, independent of this International Standard, is necessary between sender and recipient.

6.3.6 Surrogate code points

Code points D800 to DFFF are reserved for the use of the UTF-16 encoding form (see 9.3). The first half (D800 to DBFF) contains the high-surrogate code points and the second half (DC00 to DFFF) contains the low-surrogate code points.

6.3.7 Noncharacter code points

The status of noncharacter code points cannot be changed by future amendments. Noncharacters consist of FDD0-FDEF and any code point ending in the value FFFE or FFFF.

NOTE – Code point FFFE is reserved for “signature”. Code points FDD0 to FDEF, and FFFF can be used for internal processing uses requiring numeric values which are guaranteed not to be coded characters, such as in terminating tables, or signaling end-of-text. Furthermore, since FFFF is the largest BMP value, it may also be used as the final value in binary or sequential searching index within the context of UTF-16.

6.3.8 Reserved code points

Reserved code points are reserved for future standardization and shall not be used for any other purpose. Future editions of this International Standard may allocate characters to some of these reserved code points.

6.4 Naming of characters

This International Standard assigns a unique name to each graphic and format character. The name of a character either

- a) denotes the customary meaning of the character, or
- b) describes the shape of the corresponding graphic symbol, or
- c) follows the rule given in 24.6 for Chinese /Japanese/Korean (CJK) ideographs, or
- d) follows the rule given in 24.7 for Hangul syllables.

Some characters may have one or more alternate names called character name aliases which are correction of the original names. Additional rules to be used for constructing the names of characters are given in 24.

The list of character names, except for CJK unified ideographs and Hangul syllables, is provided in 31.

NOTE – The list of character names is also part of the Unicode character Database in:
<http://www.unicode.org/Public/UNIDATA/NamesList.txt> with the syntax described in:
<http://www.unicode.org/Public/UNIDATA/NamesList.html>.

6.5 Short identifiers for code points (UIDs)

This International Standard defines short identifiers for each code point, including code points that are reserved (unassigned). A short identifier for any code point is distinct from a short identifier for any other code point. If a character is allocated at a code point, a short identifier for that code point can be used to refer to the character allocated at that code point.

NOTE 1 – For instance, U+DC00 identifies a surrogate code point, and U+FFFF identifies a noncharacter code point. U+0025 identifies a graphic code point to which a graphic character is allocated; U+0025 also identifies that character (named PERCENT SIGN).

NOTE 2 – These short identifiers are independent of the language in which this standard is written, and are thus retained in all translations of the text.

The following alternative forms of notation of a short identifier are defined here.

- a) The six-digit form of short identifier consists of the sequence of six hexadecimal digits that represents the code point of the character (see 6.2).
- b) The four-to-five-digit form of short identifier shall consist of the last four to five digits of the six-digit form. Leading zeroes beyond four digits are suppressed.
- c) The character “+” (PLUS SIGN) may, as an option, precede the digit form of short identifier.
- d) The prefix letter “U” (LATIN CAPITAL LETTER U) may, as an option, precede any of the three forms of short identifier defined in a) to c) above.

The capital letters A to F, and U that appear within short identifiers may be replaced by the corresponding small letters.

The full syntax of the notation of a short identifier, in Backus-Naur form, is

$$\{ U \mid u \} \{ + \} (xxxx \mid xxxxx \mid xxxxxx)$$

where “x” represents one hexadecimal digit (0 to 9, A to F, or a to f).

EXAMPLE

The short identifier for LATIN SMALL LETTER LONG S may be notated in any of the following forms:

017F +017F U017F U+017F

Any of the capital letters may be replaced by the corresponding small letter.

6.6 UCS Sequence Identifiers

This International Standard defines an identifier for any sequence of code points taken from the standard. Such an identifier is known as a UCS Sequence Identifier (USI). For a sequence of n code points it has the following form:

$$\langle \text{UID1, UID2, ..., UIDn} \rangle$$

where UID1, UID2, etc. represent the short identifiers of the corresponding code points, in the same order as those code points appear in the sequence. If each of the code points in such a sequence has a character allocated to it, the USI can be used to identify the sequence of characters allocated at those code points. The syntax for UID1, UID2, etc. is specified in 6.5. A COMMA character (optionally followed by a SPACE character) separates the UIDs. The UCS Sequence Identifier includes at least two UIDs; it begins with a LESS-THAN SIGN and is terminated by a GREATER-THAN SIGN.

The full syntax of the notation of a UCS Sequence Identifier, in Backus-Naur form, is

$$\langle ((xxxx \mid xxxxx \mid xxxxxx) ((", " \text{ space?}) (xxxx \mid xxxxx \mid xxxxxx))^+ \rangle$$

where “x” represents one hexadecimal digit (0 to 9, A to F, or a to f).

NOTE – UCS Sequences Identifiers cannot be used for specification of subset content. They may be used outside this standard to identify: composite sequences for mapping purposes, font repertoire, etc.

6.7 Octet sequence identifiers

To represent serialized octet in the context of the encoding schemes definition (see Clause 10), this International Standard defines an identifier for serialized octet sequence. For a sequence of n octets it has the following form:

$$\langle \text{XX}_1 \text{XX}_2 \dots \text{XX}_n \rangle$$

where xx_1 , xx_2 , and xx_n , represents the first, second, and n^{th} octets using two hexadecimal digits for each octet.

7 Revision and updating of the UCS

The revision and updating of this coded character set will be carried out by ISO/IEC JTC1/SC2.

The names and code points allocation of all characters in this coded character set shall remain unchanged in all future editions and amendments of this standard. This also includes character name aliases.

NOTE – Character name aliases are created to denote errors in the character names which cannot be fixed after publication of the standard.

8 Subsets

8.1 General

This International Standard provides the specification of subsets of coded graphic characters for use in interchange, by originating devices, and by receiving devices.

There are two alternatives for the specification of subsets: limited subset and selected subset. An adopted subset may comprise either of them, or a combination of the two.

8.2 Limited subset

A limited subset consists of a list of graphic characters in the specified subset. This specification allows applications and devices that were developed using other codes to interwork with this coded character set.

A claim of conformance referring to a limited subset shall list the graphic characters in the subset by the names of graphic characters or code points as defined in this International Standard.

8.3 Selected subset

A selected subset consists of a list of collections of graphic characters as defined in this International Standard. The collections from which the selection may be made are listed in Annex A. A selected subset shall always automatically include the code points from 0020 to 007E.

A claim of conformance referring to a selected subset shall list the collections chosen as defined in this International Standard.

9 UCS encoding forms

9.1 General

This International Standard provides three encoding forms expressing each UCS scalar value in a unique sequence of one or more code units. These are named UTF-8, UTF-16, and UTF-32 respectively.

9.2 UTF-8

UTF-8 is the UCS encoding form that assigns each UCS scalar value to an octet sequence of one to four octets, as specified in table 2.

- UCS characters from the BASIC LATIN collection are represented in UTF-8 in accordance with ISO/IEC 4873, i.e. single octets with values ranging from 20 to 7E.
- Control functions in code points from 0000 to 001F, and the control character in code point 007F, are represented without the padding octets specified in Clause 11, i.e. as single octets with values ranging from 00 to 1F, and 7F respectively in accordance with ISO/IEC 4873 and with the 8-bit structure of ISO/IEC 2022.
- Octet values 00 to 7F do not otherwise occur in the UTF-8 coded representation of any character. This provides compatibility with existing file-handling systems and communications sub-systems which parse code unit sequences for these octet values.
- The first octet in the UTF-8 coded representation of any character can be directly identified when a code unit sequence is examined, one octet at a time, starting from an arbitrary location. It indicates

the number of continuing octets (if any) in the multi-octet sequence that constitutes the code unit representation of that character.

Table 2 specifies the bit distribution for the UTF-8 encoding form, showing the ranges of UCS scalar values corresponding to one, two, three, and four octet sequences.

Table 2: UTF-8 Bit distribution

Scalar value	1 st octet	2 nd octet	3 rd octet	4 th octet
00000000xxxxxx	0xxxxxx			
0000yyyyxxxxxx	110yyyy	10xxxxx		
zzzzyyyyyyxxxxx	1110zzzz	10yyyyy	10xxxxx	
000uuuuuzzzzyyyyyyxxxxx	11110uuu	10uuzzzz	10yyyyy	10xxxxx

Because surrogate code points are not UCS scalar values, any UTF-8 sequence that would otherwise map to code points D800-DFFF is ill-formed.

Table 3 lists all the ranges (inclusive) of the octet sequences that are well-formed in UTF-8. Any UTF-8 sequence that does not match the patterns listed in table 3 is ill-formed.

Table 3: Well-formed UTF-8 Octet sequences

Code points	1 st octet	2 nd octet	3 rd octet	4 th octet
0000-007F	00-7F			
0080-07FF	C2-DF	80-BF		
0800-0FFF	E0	A0-BF	80-BF	
1000-CFFF	E1-EC	80-BF	80-BF	
D000-D7FF	ED	80-9F	80-BF	
E000-FFFF	EE-EF	80-BF	80-BF	
10000-3FFFF	F0	90-BF	80-BF	80-BF
40000-FFFFFF	F1-F3	80-BF	80-BF	80-BF
100000-10FFFF	F4	80-8F	80-BF	80-BF

As a consequence of the well-formedness conditions specified in table 9.2, the following octet values are disallowed in UTF-8: C0-C1, F5-FE.

9.3 UTF-16

UTF-16 is the UCS encoding form that assigns each UCS scalar value to a sequence of one to two unsigned 16-bit code units, as specified in table 4.

In the UTF-16 encoding form, code points in the range 0000-D7FF and E000-FFFF are represented as a single 16-bit code unit; code points in the range 10000-10FFFF are represented as pairs of 16-bit code units. These pairs of special code units are known as surrogate pairs.

The values of the code units used for surrogate pairs are disjoint from the code units used for the single code unit representation, thus maintaining non-overlap for all code point representations in UTF-16.

UTF-16 optimizes the representation of characters in the BMP which contains the vast majority of common use characters.

Because surrogate code points are not UCS scalar values, unpaired surrogate code units are ill-formed.

Table 4 specifies the bit distribution for the UTF-16 encoding form. Calculation of the surrogate pair values involves subtraction of 10000 hexadecimal to account for the starting offset to the scalar value (expressed as 'www = uuuu-1' in the table).

Table 4: UTF-16 Bit distribution

Scalar value	UTF-16
xxxxxxxxxxxxxxxx	xxxxxxxxxxxxxxxx
000uuuuuxxxxxxxxxxxxxx	110110wwwxxxxxx 110111xxxxxxxxxx

NOTE – Former editions of this International Standard included references to a two-octet BMP form called UCS-2 which would be a subset of the UTF-16 encoding form restricted to the BMP UCS scalar values. The UCS-2 form is deprecated.

9.4 UTF-32 (UCS-4)

UTF-32 (or UCS-4) is the UCS encoding form that assigns each UCS scalar value to a single unsigned 32-bit code unit. The terms UTF-32 and UCS-4 can be used interchangeably to designate this encoding form.

Because surrogate code points are not UCS scalar values, UTF-32 code units in the range 0000 D800-0000 DFFF are ill-formed.

10 UCS Encoding schemes

10.1 General

Encoding schemes are octet serializations specific to each UCS encoding form, including the specification of a signature, if allowed. The signature is the code unit sequence corresponding to the code point FEFF ZERO WIDTH NO-BREAK SPACE in the corresponding encoding form. When used, a signature at the beginning of a stream of serialized octets indicates the order of the octets within the encoding form used for the representation of the characters.

This International Standard specifies seven encoding schemes: UTF-8, UTF-16BE, UTF-16LE, UTF-16, UTF-32BE, UTF-32LE, and UTF-32.

10.2 UTF-8

The UTF-8 encoding scheme serializes a UTF-8 code unit sequence in exactly the same order as the code unit sequence itself.

When represented in UTF-8, the signature turns into the octet sequence <EF BB BF>. Its usage at the beginning of a UTF-8 data stream is neither required or recommended but does not affect conformance.

10.3 UTF-16BE

The UTF-16BE encoding scheme serializes a UTF-16 code unit sequence by ordering octets in a way that the more significant octet precedes the less significant octet (also known as big-endian ordering).

In UTF-16BE, an initial octet sequence of <FE FF> is interpreted as FEFF ZERO WIDTH NO-BREAK SPACE and does not convey a signature meaning.

10.4 UTF-16LE

The UTF-16LE encoding scheme serializes a UTF-16 code unit sequence by ordering octets in a way that the less significant octet precedes the more significant octet (also known as little-endian ordering).

In UTF-16LE, an initial octet sequence of <FF FE> is interpreted as FEFF ZERO WIDTH NO-BREAK SPACE and does not convey a signature meaning.

10.5 UTF-16

The UTF-16 encoding scheme serializes a UTF-16 code unit sequence by ordering octets in a way that either the less significant octet precedes or follows the more significant octet.

In the UTF-16 encoding scheme, the initial signature read as <FE FF> indicates that the more significant octet precedes the less significant octet, and <FF FE> the reverse. The signature is not part of the textual data.

In the absence of signature, the octet order of the UTF-16 encoding scheme is that the more significant octet precedes the less significant octet.

10.6 UTF-32BE

The UTF-32BE encoding scheme serializes a UTF-32 code unit sequence by ordering octets in a way that the more significant octets precede the less significant octets (also known as big-endian ordering).

In UTF-32BE, an initial octet sequence of <00 00 FE FF> is interpreted as FEFF ZERO WIDTH NO-BREAK SPACE and does not convey a signature meaning.

10.7 UTF-32LE

The UTF-32LE encoding scheme serializes a UTF-32 code unit sequence by ordering octets in a way that the less significant octets precede the more significant octets (also known as little-endian ordering).

In UTF-32LE, an initial octet sequence of <FF FE 00 00> is interpreted as FEFF ZERO WIDTH NO-BREAK SPACE and does not convey a signature meaning.

10.8 UTF-32

The UTF-32 encoding scheme serializes a UTF-32 code unit sequence by ordering octets in a way that either the less significant octets precede or follow the more significant octets.

In the UTF-32 encoding scheme, the initial signature read as <00 00 FE FF> indicates that the more significant octet precedes the less significant octet, and <FF FE 00 00> the reverse. The signature is not part of the textual data.

In the absence of signature, the octet order of the UTF-32 encoding scheme is that the more significant octets precede the less significant octets.

11 Use of control functions with the UCS

This coded character set provides for use of control functions encoded according to ISO/IEC 6429 or similarly structured standards for control functions, and standards derived from these. A set or subset of such coded control functions may be used in conjunction with this coded character set. These standards encode a control function as a sequence of one or more octets.

When a control character of ISO/IEC 6429 is used with this coded character set, its coded representation as specified in ISO/IEC 6429 shall be padded to correspond with the number of octets in code unit of the adopted encoded form (see Clause 9). Thus, the least significant octet shall be the bit combination specified in ISO/IEC 6429, and the more significant octet(s) shall be zeros.

For example, the control character FORM FEED is represented by "000C" in the UTF-16 encoding form, and "0000 000C" in the UTF-32 encoding form.

For escape sequences, control sequences, and control strings (see ISO/IEC 6429) consisting of a coded control character followed by additional bit combinations in the range 20 to 7F, each bit combination shall be padded by octet(s) with value 00.

For example, the escape sequence "ESC 02/00 04/00" is represented by "1B 20 40" in the UTF-8 encoding form, by "001B 0020 0040" in the UTF-16 encoding form, and "0000001B 00000020 00000040" in the UTF-32 encoding form.

NOTE 1 – The term "character" appears in the definition of many of the control functions specified in ISO/IEC 6429, to identify the elements on which the control functions will act. When such control functions are applied to coded characters according to this International Standard, the action of those control functions will depend on the type of element from this International Standard that has been chosen, by the application, to be the element (or character) on which the control functions act. These

ISO/IEC 10646:2014 (E)

elements may be chosen to be characters (non-combining characters and/or combining characters) or may be chosen in other ways (such as composite sequences) when applicable.

Code extension control functions for the ISO/IEC 2022 code extension techniques (such as designation escape sequences, single shift, and locking shift) shall not be used with this coded character set.

NOTE 2 – The following list provides the long names from ISO/IEC 6429 used in association with the control characters.

0000 NULL	007F DELETE
0001 START OF HEADING	0082 BREAK PERMITTED HERE
0002 START OF TEXT	0083 NO BREAK HERE
0003 END OF TEXT	0084 INDEX
0004 END OF TRANSMISSION	0085 NEXT LINE
0005 ENQUIRY	0086 START OF SELECTED AREA
0006 ACKNOWLEDGE	0087 END OF SELECTED AREA
0007 BELL	0088 CHARACTER TABULATION SET
0008 BACKSPACE	0089 CHARACTER TABULATION WITH JUSTIFICATION
0009 CHARACTER TABULATION	008A LINE TABULATION SET
000A LINE FEED	008B PARTIAL LINE FORWARD
000B LINE TABULATION	008C PARTIAL LINE BACKWARD
000C FORM FEED	008D REVERSE LINE FEED
000D CARRIAGE RETURN	008E SINGLE-SHIFT TWO
000E SHIFT-OUT	008F SINGLE-SHIFT THREE
000F SHIFT-IN	0090 DEVICE CONTROL STRING
0010 DATA LINK ESCAPE	0091 PRIVATE USE ONE
0011 DEVICE CONTROL ONE	0092 PRIVATE USE TWO
0012 DEVICE CONTROL TWO	0093 SET TRANSMIT STATE
0013 DEVICE CONTROL THREE	0094 CANCEL CHARACTER
0014 DEVICE CONTROL FOUR	0095 MESSAGE WAITING
0015 NEGATIVE ACKNOWLEDGE	0096 START OF GUARDED AREA
0016 SYNCHRONOUS IDLE	0097 END OF GUARDED AREA
0017 END OF TRANSMISSION BLOCK	0098 START OF STRING
0018 CANCEL	009A SINGLE CHARACTER INTRODUCER
0019 END OF MEDIUM	009B CONTROL SEQUENCE INTRODUCER
001A SUBSTITUTE	009C STRING TERMINATOR
001B ESCAPE	009D OPERATING SYSTEM COMMAND
001C INFORMATION SEPARATOR FOUR	009E PRIVACY MESSAGE
001D INFORMATION SEPARATOR THREE	009F APPLICATION PROGRAM COMMAND
001E INFORMATION SEPARATOR TWO	
001F INFORMATION SEPARATOR ONE	

The control character 0084 INDEX has been removed from ISO/IEC 6429:1992. In addition, the control characters 000E and 000F are named SHIFT-OUT and SHIFT-IN respectively in 7-bit environment and LOCKING-SHIFT ONE and LOCKING-SHIFT ZERO respectively in 8-bit environment.

12 Declaration of identification of features

12.1 Purpose and context of identification

Code unit sequences conforming to this International Standard are intended to form all or part of a composite unit of coded information that is interchanged between an originator and a recipient. The identification of this International Standard (including the encoding form and the encoding scheme) and any subset of the coding space that have been adopted by the originator should also be available to the recipient. The route by which such identification is communicated to the recipient is outside the scope of this International Standard.

However, some standards for interchange of coded information may permit, or require, that the coded representation of the identification applicable to the code unit sequence forms a part of the interchanged information. Clause 12 specifies a coded representation for the identification of UCS and a subset of this International Standard, and also of a C0 and a C1 set of control functions from ISO/IEC 6429 for use in conjunction with this International Standard. Such coded representations provide all or part of an identification data element, which may be included in information interchange in accordance with the relevant standard.

In the context of these identifications, because the more significant octets shall precede the less significant octets when serialized, the only encoding schemes that can be selected are UTF-8, UTF-16BE, and UTF-32BE according to the relevant encoding forms (UTF-8, UTF-16, and UTF-32 respectively).

If two or more of the identifications are present, the order of those identifications shall follow the order as specified in Clause 12.

NOTE – An alternative method of identification is described in Annex N.

12.2 Identification of a UCS encoding scheme

When the escape sequences from ISO/IEC 2022 are used, the identification of a UCS encoding scheme (see Clause 9) specified by this International Standard shall be by a designation sequence chosen from the following list:

ESC 02/05 02/15 04/09

UTF-8 encoding form; UTF-8 encoding scheme

ESC 02/05 02/15 04/12

UTF-16 encoding form; UTF-16BE encoding scheme

ESC 02/05 02/15 04/06

UTF-32 encoding form; UTF-32BE encoding scheme

NOTE – The following designation sequences: ESC 02/05 02/15 04/00, ESC 02/05 02/15 04/01, ESC 02/05 02/15 04/03, ESC 02/05 02/15 04/04, ESC 02/05 02/15 04/07, ESC 02/05 02/15 04/08, ESC 02/05 02/15 04/10, ESC 02/05 02/15 04/11 used in previous versions of this standard to identify implementation levels 1 and 2 are deprecated. The remaining designation sequences correspond to the former level 3 which is now the only supported content definition for code unit sequences.

ESC 02/05 04/07

UTF-8 encoding form; UTF-8 encoding scheme

If such an escape sequence appears within a code unit sequence conforming to ISO/IEC 2022, it shall consist only of the sequences of bit combinations as shown above.

If such an escape sequence appears within a code unit sequence conforming to this International Standard, it shall be padded in accordance with Clause 11 when the identified encoding form is either UTF-16 or UTF-32. No padding is necessary when the identified encoding form is UTF-8. See also 12.5.

12.3 Identification of subsets of graphic characters

When the control sequences of ISO/IEC 6429 are used, the identification of subsets (see Clause 8) specified by this International Standard shall be by a control sequence IDENTIFY UNIVERSAL CHARACTER SUBSET (IUCS) as shown below.

CSI Ps... 02/00 06/13

Ps... means that there can be any number of selective parameters. The parameters are to be taken from the subset collection numbers as shown in 9. When there is more than one parameter, each parameter value is separated by an octet with value 03/11.

Parameter values are represented by digits where octet values 03/00 to 03/09 represent digits 0 to 9.

If such an escape sequence appears within a code unit sequence conforming to ISO/IEC 2022, it shall consist only of the sequences of bit combinations as shown above.

If such a control sequence appears within a code unit sequence conforming to this International Standard, it shall be padded in accordance with Clause 11.

12.4 Identification of control function set

When the escape sequences from ISO/IEC 2022 are used, the identification of each set of control functions (see Clause 11) of ISO/IEC 6429 to be used in conjunction with ISO/IEC 10646 shall be an identifier sequence of the type shown below.

ESC 02/01 04/00 identifies the full C0 set of ISO/IEC 6429

ESC 02/02 04/03 identifies the full C1 set of ISO/IEC 6429

ISO/IEC 10646:2014 (E)

For other C0 or C1 sets, the final octet F shall be obtained from the International Register of Coded Character Sets. The identifier sequences for these sets shall be

ESC 02/01 F	identifies a C0 set
ESC 02/02 F	identifies a C1 set

If such an escape sequence appears within a code unit sequence conforming to ISO/IEC 2022, it shall consist only of the sequences of bit combinations as shown above.

If such an escape sequence appears within a code unit sequence conforming to this International Standard, it shall be padded in accordance with Clause 11.

12.5 Identification of the coding system of ISO/IEC 2022

When the escape sequences from ISO/IEC 2022 are used, the identification of a return, or transfer, from UCS to the coding system of ISO/IEC 2022 shall be by the escape sequence ESC 02/05 04/00. If such an escape sequence appears within a code unit sequence conforming to this International Standard, it shall be padded in accordance with Clause 11.

If such an escape sequence appears within a code unit sequence conforming to ISO/IEC 2022, it shall consist only of the sequence of bit combinations as shown above.

NOTE – Escape sequence ESC 02/05 04/00 is normally used for return to the restored state of ISO/IEC 2022. The escape sequence ESC 02/05 04/00 specified here is sometimes not exactly as specified in ISO/IEC 2022 due to the presence of padding octets. For this reason the escape sequences used in 12.2 for the identification of UCS (except for ESC 02/05 04/07) include the octet 02/15 to indicate that the return does not always conform to that standard.

13 Structure of the code charts and lists

Clause 31 sets out the detailed code charts and the lists of character names for the graphic characters. It specifies for each character their graphic symbol, their coded representation, and their character name.

NOTE – Clause 31 also includes additional information on characters clarifying some feature of a character, such as its naming or usage, or its associated graphic symbol.

Graphic symbols are to be regarded as typical visual representations of the corresponding graphic characters. This International Standard does not attempt to prescribe the exact shape or colour of each character. The shape is affected by the design of the font or other representation method employed, which is out of scope. Although the representative glyphs in this International Standard are consistently presented in black and white, it does not prevent implementations from using graphic symbols with some specific colour or even with multiple colours, fully or partly animated graphics, or both. When characters are typically associated with a particular colour, conventions of European heraldry are used to represent those colours in monochromatic line drawings in the code charts. Furthermore, the usage of 'BLACK' and 'WHITE' in character names does not imply a specific colour. It is simply a distinction between a filled character and an outline character.

Graphic characters specified in this International Standard are uniquely identified by their names. This does not imply that the graphic symbols by which they are commonly imaged are always different. Examples of graphic characters with similar graphic symbols are LATIN CAPITAL LETTER A, GREEK CAPITAL LETTER ALPHA and CYRILLIC CAPITAL LETTER A.

The meaning attributed to any character is not specified by this International Standard; it may differ from country to country, or from one application to another.

For the alphabetic scripts, the general principle has been to arrange the characters within any row in approximate alphabetic sequence; where the script has capital and small letters, these are arranged in pairs. However, this general principle has been overridden in some cases. For example, for those scripts for which a relevant standard exists, the characters are allocated according to that standard. This arrangement within the code charts will aid conversion between the existing standards and this coded character set. In general, however, it is anticipated that conversion between this coded character set and any other coded character set will use a table lookup technique.

It is not intended, nor will it often be the case, that the characters needed by any one user will be found all grouped together in one part of the code charts.

Furthermore, the user of any script will find that needed characters may have been coded elsewhere in this coded character set. This especially applies to the digits, to the symbols, and to the use of Latin letters in dual-script applications.

Therefore, in using this coded character set, the reader is advised to refer first to the block names list in Annex A.2 or an overview of the Planes in figures 7 to 12, and then to turn to the specific code chart for the relevant script and for symbols and digits. In addition, Annex G contains an alphabetically sorted list of character names.

14 Block and collection names

14.1 Block names

Named blocks of contiguous code points are specified within a plane for the purpose of allocation of characters sharing some common characteristic, such as script. The blocks specified within the BMP, SMP, SIP and SSP are listed in A.2, and are illustrated in figures 7 to 12.

Rules to be used for constructing the names of blocks are given in 24.5.1.

14.2 Collection names

Collections are shown in Annex A.

Rules to be used for constructing the names of collections are given in 24.5.2.

15 Mirrored characters in bidirectional context

15.1 Mirrored characters

A class of characters has special significance in the context of bidirectional text. The interpretation and rendering of any of these characters depend on the direction of the character being rendered that is in effect at the point in the code unit sequence where the coded representation of the character appears. The list of these characters is determined by having the 'Bidi_Mirrored' property set to 'Y' in the Unicode Standard. These values shall be determined according to the Unicode Standard Bidi Mirrored property (see Clause 3).

NOTE 1 – Typically, a mirrored character has its image mirrored horizontally in text that is laid out from right to left. However, for some mathematical symbols, the 'mirrored' form is not an exact mirror image. See the Unicode Technical Report #25, "Unicode Support for Mathematics" for additional details.

This character mirroring is not limited to paired characters and shall be applied to all characters belonging to that class.

EXAMPLE

In a right-to-left text segment, the GREATER-THAN SIGN (rendered as ">" in left-to-right text) may be rendered as the "<" graphic symbol.

NOTE 2 – Many ancient scripts and some scripts in modern use can be written either right-to-left or left-to-right. It is often customary for one of these scripts to use the appropriately mirrored graphical symbol for any character represented by a graphic symbol that is not symmetric around the vertical axis. In such cases, it is up to the rendering system to display the graphic image appropriate for the writing direction employed. The directionality of the representative graphic symbol shown in the character code charts matches the default writing direction for the script. Characters belonging to these scripts have the "Bidi_Mirrored" property set to 'N' in the Unicode Standard (see reference to the Unicode Standard Bidi Mirrored property in 3).

Examples of such scripts include, but are not limited to, Old Italic, an ancient script for which the default writing direction in this standard is left-to-right, and Cypriot, an ancient script for which the default writing direction in this standard is right-to-left.

15.2 Directionality of bidirectional text

The Unicode Bidirectional Algorithm (see Clause 3) describes the algorithm used to determine the directionality for bidirectional text. It shall be used in the context of this International Standard.

16 Special characters

16.1 General

There are some characters that do not have printable graphic symbols or are otherwise special in some ways.

16.2 Space characters

The following characters are space characters. They represent all characters which have the General Category value set to 'Zs'.

Code Point	Name		
0020	SPACE	2005	FOUR-PER-EM SPACE
00A0	NO-BREAK SPACE	2006	SIX-PER-EM SPACE
1680	OGHAM SPACE MARK	2007	FIGURE SPACE
2000	EN QUAD	2008	PUNCTUATION SPACE
2001	EM QUAD	2009	THIN SPACE
2002	EN SPACE	200A	HAIR SPACE
2003	EM SPACE	202F	NARROW NO-BREAK SPACE
2004	THREE-PER-EM SPACE	205F	MEDIUM MATHEMATICAL SPACE
		3000	IDEOGRAPHIC SPACE

NOTE 1 – The character 1680 OGHAM SPACE MARK is typically represented with a visible glyph showing the central stem-line, and it is only represented by a blank glyph in a "stemless" style font.

NOTE 2 – The character 202F NARROW NO-BREAK-SPACE is a non-breaking space. It is similar to 00A0 NO-BREAK SPACE, except that it is rendered with a narrower width. When used with the Mongolian script this character is usually rendered at one-third of the width of a normal space, and it separates a suffix from the Mongolian word-stem. This allows for the normal rules of Mongolian character shaping to apply, while indicating that there is no word boundary at that position.

16.3 Currency symbols

Currency symbols in this International Standard do not necessarily identify the currency of a country. For example, YEN SIGN can be used for Japanese Yen and Chinese Yuan. Also, DOLLAR SIGN is used in numerous countries including the United States of America.

16.4 Format characters

The following characters are format characters (see 6.3.3). They represent all characters which have the General Category value set to 'Cf', 'Zl', and 'Zp'. See Annex F.

Code Point	Name		
00AD	SOFT HYPHEN	2063	INVISIBLE SEPARATOR
0600	ARABIC NUMBER SIGN	2064	INVISIBLE PLUS
0601	ARABIC SIGN SANAH	2066	LEFT-TO-RIGHT ISOLATE
0602	ARABIC FOOTNOTE MARKER	2067	RIGHT-TO-LEFT ISOLATE
0603	ARABIC SIGN SAFHA	2068	FIRST STRONG ISOLATE
0604	ARABIC SIGN SAMVAT	2069	POP DIRECTIONAL ISOLATE
0605	ARABIC NUMBER MARK ABOVE	206A	INHIBIT SYMMETRIC SWAPPING
061C	ARABIC LETTER MARK	206B	ACTIVATE SYMMETRIC SWAPPING
06DD	ARABIC END OF AYAH	206C	INHIBIT ARABIC FORM SHAPING
070F	SYRIAC ABBREVIATION MARK	206D	ACTIVATE ARABIC FORM SHAPING
180E	MONGOLIAN VOWEL SEPARATOR	206E	NATIONAL DIGIT SHAPES
200B	ZERO WIDTH SPACE	206F	NOMINAL DIGIT SHAPES
200C	ZERO WIDTH NON-JOINER	FEFF	ZERO WIDTH NO-BREAK SPACE
200D	ZERO WIDTH JOINER	FFF9	INTERLINEAR ANNOTATION ANCHOR
200E	LEFT-TO-RIGHT MARK	FFFA	INTERLINEAR ANNOTATION SEPARATOR
200F	RIGHT-TO-LEFT MARK	FFFB	INTERLINEAR ANNOTATION TERMINATOR
2028	LINE SEPARATOR	110BD	KAITHI NUMBER SIGN
2029	PARAGRAPH SEPARATOR	1BCA0	SHORTHAND FORMAT LETTER OVERLAP
202A	LEFT-TO-RIGHT EMBEDDING	1BCA1	SHORTHAND FORMAT CONTINUING OVERLAP
202B	RIGHT-TO-LEFT EMBEDDING	1BCA2	SHORTHAND FORMAT DOWN STEP
202C	POP DIRECTIONAL FORMATTING	1BCA3	SHORTHAND FORMAT UP STEP
202D	LEFT-TO-RIGHT OVERRIDE	1D173	MUSICAL SYMBOL BEGIN BEAM
202E	RIGHT-TO-LEFT OVERRIDE	1D174	MUSICAL SYMBOL END BEAM
2060	WORD JOINER	1D175	MUSICAL SYMBOL BEGIN TIE
2061	FUNCTION APPLICATION	1D176	MUSICAL SYMBOL END TIE
2062	INVISIBLE TIMES	1D177	MUSICAL SYMBOL BEGIN SLUR

1D178	MUSICAL SYMBOL END SLUR	E0001	LANGUAGE TAG
1D179	MUSICAL SYMBOL BEGIN PHRASE	E0020-E007F	TAG SPACE to CANCEL TAG
1D17A	MUSICAL SYMBOL END PHRASE		

16.5 Ideographic description characters

An Ideographic Description Character (IDC) is a graphic character, which is used with a sequence of other graphic characters to form an Ideographic Description Sequence (IDS). Such a sequence may be used to describe an ideographic character which is not specified with this International Standard. The Annex I describes them in more details. The list of IDC follows:

Code Point	Name
2FF0	IDEOGRAPHIC DESCRIPTION CHARACTER LEFT TO RIGHT
2FF1	IDEOGRAPHIC DESCRIPTION CHARACTER ABOVE TO BELOW
2FF2	IDEOGRAPHIC DESCRIPTION CHARACTER LEFT TO MIDDLE AND RIGHT
2FF3	IDEOGRAPHIC DESCRIPTION CHARACTER ABOVE TO MIDDLE AND BELOW
2FF4	IDEOGRAPHIC DESCRIPTION CHARACTER FULL SURROUND
2FF5	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM ABOVE
2FF6	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM BELOW
2FF7	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LEFT
2FF8	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM UPPER LEFT
2FF9	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM UPPER RIGHT
2FFA	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LOWER LEFT
2FFB	IDEOGRAPHIC DESCRIPTION CHARACTER OVERLAID

16.6 Variation selectors and variation sequences

16.6.1 General

Variation selectors are a specific class of combining characters immediately following a base character for which there is no canonical mapping and no equivalent composite sequence. The purpose is to indicate a specific variant form of graphic symbol for that base character. The character sequence consisting of such base character followed by a variation selector is called a variation sequence.

NOTE – The variation selector only selects a specific *appearance* among those acceptable for an already encoded character. It is not intended as a general code extension mechanism.

Variation selectors are made of all coded characters included in the blocks VARIATION SELECTORS and VARIATION SELECTORS SUPPLEMENT and the three Mongolian Free Variation Selectors (FVS1 To FVS3).

A variation sequence whose base character is a CJK unified ideograph and whose variation selector is from the VARIATION SELECTORS SUPPLEMENT BLOCK is called an ideographic variation sequence. All other variation sequences are called standardized variation sequences. The variant form of a graphic symbol specified by a standardized variation sequence is called a standardized variant.

Only the variation sequences defined or referenced in 16.6 indicate a specific variant form of graphic symbol; all other such sequences are undefined. Furthermore, variation selectors following other base characters and any non-base characters have no effect on the selection of the graphic symbol for that character.

16.6.2 Standardized variation sequences

Standardized variation sequences are defined by a machine-readable format that is accessible as a link.

The content linked to is a plain text file, using ISO/IEC 646-IRV characters with CARRIAGE RETURN/LINE FEED as end of line mark that specifies Standardized Variations Sequences. Each line in the text file contains the following information organized in two or three fields:

- 1st field: Variation sequence expressed as a UC S Sequence Identifier (using a modified syntax for the USI, omitting commas and angle brackets),
- 2nd field: Description of the variation sequence,
- 3rd field (optional): Shaping environments where the variation sequence applies. The possible values are: isolate, initial, medial, final.

The fields are delimited by a SEMICOLON (;) followed optionally by zero or more SPACE characters. The last field may be followed by a comment starting with a NUMBER SIGN (#) describing the name of the base character. Comment lines, starting with a NUMBER SIGN (#) are informational only. Comment lines and blank lines in the text file should be ignored by any automatic process which parses the data file to extract the normative list of Standardized Variants.

[Click on this highlighted text to access the reference file.](#)

NOTE 1 – The content is also available as a separate viewable file in the same file directory as this document. The file is named “UCSVariants.txt”. These sequences are also described as *StandardizedVariants.html* in the Unicode character database (<http://www.unicode.org/Public/UNIDATA/StandardizedVariants.html>).

The Standardized Variations Sequences contain sequences associated with allowed base characters from the following categories:

- Mathematical symbols

NOTE 2 – The VARIATION SELECTOR-1 (FE00) is the only variation selector used with mathematical symbols.

- Mongolian characters. Only some presentation forms of the base Mongolian characters used with the Mongolian free variation selectors produce variant appearances.

NOTE 3 – The Mongolian characters have various presentation forms depending on their position in a code unit sequence. These presentations forms are called isolate, initial, medial and final.

- Manichaeen characters

NOTE 4 – These are similar to Mongolian variations sequences in depending on their position in a code unit sequence, but using only isolate and final presentation forms.

- Phags-pa characters. These variation sequences do not select fixed visual representation; rather, they select a representation that is reversed from the normal form predicted by the preceding character.
- Pictographic symbols. The range of presentations may include a traditional black and white text style, using FE0E VARIATION SELECTOR-15, or an ‘emoji’ style, using FE0F VARIATION SELECTOR-16, whose presentation often involves color/grayscale and/or animation.

NOTE 5 – Standardized variations sequences involving the characters 0023 NUMBER SIGN and the range 0030-0039 (DIGIT ZERO to DIGIT NINE) are intended for use with 20E3 COMBINING ENCLOSING KEYCAP, such as in:

<0023, FE0E, 20E3> NUMBER SIGN inside a COMBINING ENCLOSING KEYCAP in text style.

- CJK Unified Ideographs. Each of these variation sequences corresponds to a CJK compatibility ideograph. Its specified appearance is that of the corresponding CJK compatibility ideograph.

NOTE 6 – All normalization forms replace CJK compatibility ideographs with the corresponding CJK unified ideographs, but leave the variation sequences unchanged (see Clause 21). In contexts where normalization forms are used and the distinction between the CJK compatibility ideographs and CJK unified ideographs is desired, the usage of variation sequences is a mechanism to maintain that distinction. No equivalence between these variation sequences and the corresponding compatibility ideographs is defined. Conversion considerations are out of scope of this International Standard

16.6.3 Ideographic variation sequences

Variations sequences composed of a unified ideograph as the base character and one of VARIATION SELECTOR-17 to VARIATION SELECTOR-256 from the Supplementary Special-purpose Plane (SSP) are registered in the Ideographic Variation Database defined by Unicode Technical Standard #37. The version referenced in Clause 3 shall be used in the context of this International Standard.

NOTE – This International Standard incorporates by reference the variation sequences listed in version 2010-11-14 of the Ideographic Variation Database, as described at <http://www.unicode.org/ivd/data/2012-03-02/>.

17 Presentation forms of characters

Each presentation form of a character provides an alternative form, for use in a particular context, to the nominal form of the character or sequence of characters from the other zones of graphic characters. The transformation from the nominal form to the presentation forms may involve substitution, superimposition, or combination.

The rules for the superimposition, choice of differently shaped characters, or combination into ligatures, or conjuncts, which are often of extreme complexity, are not specified in this International Standard.

In general, presentation forms are not intended to be used as a substitute for the nominal forms of the graphic characters specified elsewhere within this coded character set. However, specific applications may encode these presentation forms instead of the nominal forms for specific reasons among which is compatibility with existing devices. The rules for searching, sorting, and other processing operations on presentation forms are outside the scope of this International Standard.

Within the BMP these characters are mostly allocated to code points within rows from FB to FF.

18 Compatibility characters

Compatibility characters are included in this International Standard primarily for compatibility with existing coded character sets to allow two-way code conversion without loss of information.

Within the BMP many of these characters are allocated to code points within rows F9, FA, FE, and FF, and within rows 31 and 33. Some compatibility characters are also allocated within other rows.

NOTE 1 – There are twelve code points in the row FA of the BMP which are allocated to CJK Unified Ideographs.

Within the Supplementary Ideographic Plane (SIP) these characters are allocated to code points within rows F8 to FA.

The CJK compatibility ideographs are ideographs that should have been unified with one of the CJK unified ideographs, per the unification rule described in Annex S. However, they are included in this International Standard as separate characters, because, based on various national, cultural, or historical reasons for some specific country and region, some national and regional standards assign separate code points for them.

NOTE 2 – For this reason, compatibility ideographs should only be used for maintaining and guaranteeing a round trip conversion with the specific national, regional, or other standard. Other usage is strongly discouraged.

NOTE 3 – Because compatibility ideographs are not preserved through any normalization forms, use of standardized variation sequences for CJK Unified Ideographs (see 16.6) may be preferred in contexts where normalization forms are used and the distinction between CJK compatibility ideographs and the corresponding CJK Unified ideographs needs to be maintained. In context where compatibility ideographs should be preserved, normalization forms cannot be used.

19 Order of characters

Usually, coded characters appear in a code unit sequence in logical order (logical or backing store order corresponds approximately to the order in which characters are entered from the keyboard, after corrections such as insertions, deletions, and overtyping have taken place). This applies even when characters of different dominant direction are mixed: left-to-right (Greek, Latin, Thai) with right-to-left (Arabic, Hebrew), or with vertical (Mongolian) script.

Some characters may not appear linearly in final rendered text. For example, the medial form of DEVANAGARI VOWEL SIGN I is displayed before the character that it logically follows in the code unit sequence.

20 Combining characters

20.1 Order of combining characters

Coded representations of combining characters shall follow that of the graphic character with which they are associated (for example, coded representations of LATIN SMALL LETTER A followed by COMBINING TILDE represent a composite sequence for Latin “ã”).

If a combining character is to be regarded as a composite sequence in its own right, it shall be coded as a composite sequence by association with the character 00A0 NO-BREAK SPACE. For example, grave accent can be composed as 00A0 NO-BREAK SPACE followed by 0300 COMBINING GRAVE ACCENT.

NOTE – Indic combining marks for vowels form a special category of combining characters, since the presentation can depend on more than one of the surrounding characters. Thus it might not be desirable to associate these Indic combining marks with the character NO-BREAK SPACE.

20.2 Combining class and canonical ordering

Each combining character has a combining class value determined by the Unicode Standard. These values shall be determined according to the Unicode Standard Combining Class property (see Clause 3). The combining class is used to determine the canonical ordering which is part of the normalization process (see Clause 21). Canonical ordering consists in ordering combining characters in the increasing order of their combining class. Combining characters with a combining class value of zero are not re-ordered relative to other characters.

20.3 Appearance in code charts

Combining characters intended to be positioned relative to the associated character are depicted within the character code charts above, below, to the right of, to the left of, in, around, or through a dotted circle to show their position relative to the base character. In presentation, these characters are intended to be positioned relative to the preceding base character in some manner, and not to stand alone or function as base characters. This is the motivation for the term “combining”.

NOTE – Diacritics are the principal class of combining characters used in European alphabets. For many other scripts used in India and South East Asia, combining characters encode vowel letters; as such they are not generally referred to as “diacritical marks”.

20.4 Alternate coded representations

Alternate coded representations of text are generated by using multiple combining characters in different orders, or using various equivalent combinations of characters and composite sequences. These alternate coded representations result in multiple representations of the same text. Normalizing (see Clause 21) these coded representations reduces significantly, but does not eliminate, the occurrences of these multiple representations.

NOTE – For example, the French word “là” may be represented by the characters LATIN SMALL LETTER L followed by LATIN SMALL LETTER A WITH GRAVE, or may be represented by the characters LATIN SMALL LETTER L followed by LATIN SMALL LETTER A followed by COMBINING GRAVE ACCENT. When the normalization forms are applied on those alternate coded representations, only one representation remains. The form of the remaining representation depends on the normalization form used.

20.5 Multiple combining characters

There are instances where more than one combining character is applied to a single graphic character. This International Standard does not restrict the number of combining characters that can follow a base character. The following rules apply to the presentation of these characters:

- a) If the combining characters can interact in presentation (for example, COMBINING MACRON and COMBINING DIAERESIS), then the position of the combining characters in the resulting graphic display is determined by the order of the coded representation of the combining characters. The presentations of combining characters are to be positioned from the base character outward. For example, combining characters placed above a base character are stacked vertically, starting with the first encountered in the sequence of coded representations and continuing for as many marks above as are required by the coded combining characters following the coded base character. For combining characters placed below a base character, the situation is inverted, with the combining characters starting from the base character and stacking downward.

An example of multiple combining characters above the base character is found in Thai, where a consonant letter can have above it one of the vowels 0E34 to 0E37 and, above that, one of four tone marks 0E48 to 0E4B. The order of the coded representation is: base consonant, followed by a vowel, followed by a tone mark.

- b) Some specific combining characters override the default stacking behaviour by being positioned horizontally rather than stacking, or by forming a ligature with an adjacent combining character. When positioned horizontally, the order of coded representations is reflected by positioning in the dominant order of the script with which they are used. For example, horizontal accents in a left-to-right script are coded left-to-right.

Prominent characters that show such override behaviour are associated with specific scripts or alphabets. For example, the COMBINING GREEK KORONIS (0343) requires that, together with a following acute or grave accent, they be rendered side-by-side above a letter, rather than the accent marks being stacked above the COMBINING GREEK KORONIS. The order of the coded representations is: the letter itself, followed by that of the breathing mark, followed by that of the accent marks. Two Vietnamese tone marks, which have the same graphic appearance as the Latin acute and grave accent marks, do not stack above the three Vietnamese vowel letters which already contain the circumflex diacritic (â, ê, ô). Instead, they form ligatures with the circumflex component of the vowel letters.

- c) If the combining characters do not interact in presentation (for example, when one combining character is above a graphic character and another is below), the resultant graphic symbol from the base character and combining characters in different orders may appear the same. For example, the coded representations of LATIN SMALL LETTER A, followed by COMBINING CARON, followed by COMBINING OGONEK may result in the same graphic symbol as the coded representations of LATIN SMALL LETTER A, followed by COMBINING OGONEK, followed by COMBINING CARON.

Combining characters in Hebrew or Arabic scripts do not normally interact. Therefore, the sequence of their coded representations in a composite sequence does not affect its graphic symbol. The rules for forming the combined graphic symbol are beyond the scope of this International Standard.

20.6 Collections containing combining characters

In some collections of characters listed in Annex A, such as collections 14 (BASIC ARABIC) or 25 (THAI), both combining characters and non-combining characters are included.

Other collections of characters listed in Annex A comprise only combining characters, for example collection 7 (COMBINING DIACRITICAL MARKS).

20.7 Combining Grapheme Joiner

The character 034F COMBINING GRAPHEME JOINER is used to indicate that adjacent characters are to be treated as a unit for the purpose of language-sensitive collation and searching. In language-sensitive collation and searching, the combining grapheme joiner shall be ignored unless it specifically occurs with a tailored collation element mapping. For rendering, the combining grapheme joiner is invisible.

NOTE – The combining grapheme joiner may be used to differentiate two usages of a combining character by using it for one of the two cases. For example, where a distinction is needed between the German umlaut and the tréma, the COMBINING GRAPHEME JOINER (034F) followed by the COMBINING DIAERESIS (0308) should be used to represent the tréma while the COMBINING DIAERESIS (0308) alone should be used to represent the German umlaut.

21 Normalization forms

Normalization forms are the mechanisms allowing the selection of a unique coded representation among alternative, but equivalent coded text representations of the same text. Normalization forms for use with this International Standard are specified in the Unicode Standard UAX#15 (see Clause 3) and shall be used in the context of this International Standard. There are four normalization forms:

- a) Normalization Form D (NFD),
- b) Normalization Form C (NFC),
- c) Normalization Form KD (NFKD),
- d) Normalization Form KC (NFKC).

NOTE 1 – The result of applying any of these normalization forms onto a code unit sequence is intended to stay stable over time. It means that the normalized representation of a code unit sequence consisting of characters assigned in this version of the standard remains normalized even when the standard is amended.

NOTE 2 – Some normalization forms favour composite sequences over shorter representations of text, others favour the shorter representations. The backward compatibility requirement is provided by establishing ISO/IEC 10646-1:2000 (2nd Edition) and ISO/IEC 10646-2:2001 (1st Edition) as the reference versions for the definition of the shorter representation of text. The union of their repertoire is identical to the fixed collection UNICODE 3.2 (see A.6.3).

NOTE 3 – The goal of normalization is to provide a unique normalized result for any given code unit sequence to facilitate, among other things, identity matching. A normalized form does not necessarily represent the optimal sequence from a linguistic point of view.

NOTE 4 – In all four normalization forms, CJK Compatibility Ideographs are replaced with the corresponding CJK Unified Ideographs. Normalization, however, does not alter variation selectors, and variation sequences are preserved. Because of this, the use of standardized variation sequences for CJK Unified Ideographs over the CJK Compatibility Ideographs is preferred in the context of normalization (see 16.6).

22 Special features of individual scripts and symbol repertoires

22.1 Hangul syllable composition method

In rendering, a sequence of Hangul Jamo (from HANGUL JAMO block: 1100 to 11FF, HANGUL JAMO EXTENDED-A block: A960 to A97F, and HANGUL JAMO EXTENDED-B block: D7B0 to D7FF) is displayed as a series of syllable blocks. Jamo can be classified into three classes: Choseong (syllable-initial characters or initial consonants), Jungseong (syllable-peak characters or medial vowels), and Jongseong (syllable-final characters or final consonants). A complete syllable block is composed of a Choseong and a Jungseong, and optionally a Jongseong.

An incomplete syllable is a string of one or more characters which does not constitute a complete syllable (for example, a Choseong alone, a Jungseong alone, a Jongseong alone, or a Jungseong followed by a Jongseong). An incomplete syllable which starts with a Jungseong shall be preceded by a CHOSEONG FILLER (115F). An incomplete syllable composed of a Jongseong alone shall be preceded by a CHOSEONG FILLER (115F) and JUNGSEONG FILLER (1160). An incomplete syllable composed of a Choseong alone shall be followed by a JUNGSEONG FILLER (1160).

NOTE 1 – Hangul Jamo are not combining characters.

NOTE 2 – When a combining character such as HANGUL SINGLE DOT TONE MARK (302E) is intended to apply to a sequence of Hangul Jamo it should be placed at the end of the sequence, after the Hangul Jamo character which completes the syllable block.

NOTE 3 – Hangul text can be represented in several different ways in this standard. Korean Standard KS X 1026-1: Information Technology - Universal Multiple-Octet Coded Character set (UCS) - Hangul - Part 1, Hangul processing guide for information interchange, provides guidelines on how to ensure interoperability in information interchange.

22.2 Features of scripts used in India and some other South Asian countries

In the code charts for Rows 09 to 0D and 0F, and for the MYANMAR block in Row 10, of the BMP (see Clause 31) the graphic symbols shown for some characters appear to be formed as compounds of the graphic symbols for two other characters in the same table.

EXAMPLE 1 Row 09 Devanagari

The graphic symbol for 0906 DEVANAGARI LETTER AA appears as if it is constructed from the graphic symbols for 0905 DEVANAGARI LETTER A and 093E DEVANAGARI VOWEL SIGN AA

EXAMPLE 2 Row 0D Malayalam

The graphic symbol for 0D08 MALAYALAM LETTER II appears as if it is constructed from the graphic symbols for 0D07 MALAYALAM LETTER I and 0D57 MALAYALAM AU LENGTH MARK

In such cases a single coded character may appear to the user to be equivalent to the sequence of two coded characters whose graphic symbols, when combined, are visually similar to the graphic symbol of that single character, as in a composite sequence (see 4.17).

A “unique-spelling” rule is defined as follows. According to this rule, no coded character from a table for Rows 09 to 0D or 0F, or for the MYANMAR block in Row 10, with the list of exceptions mentioned below, shall be regarded as equivalent to a sequence of two or more other coded characters taken from the same table.

- Two-part dependent vowel signs,
- Independent vowel 1025 MYANMAR LETTER UU,
- Consonants including a nukta sign.

NOTE – All these characters have canonical mapping consisting of a sequence of two characters.

22.3 Byzantine musical symbols

The Byzantine Musical Notation System makes use of the so-called ‘three-stripe’ effect. There are signs that appear in the Upper, Middle or Lower stripes. Other signs are known as musical characters and appear in the textual part of the notation system. Multiple signs can be stacked together in their appropriate stripe.

22.4 Source references for pictographic symbols

Some symbols are associated by reference with various industrial sources. Unlike CJK Unified ideographs, these references do not establish character identity. These sources may be associated with either single code points or sequences of code points.

The symbol sources are:

- DoCoMo Shift-JIS code
- KDDI Shift-JIS code
- SoftBank Shift-JIS code

The content linked to is a plain text file, using ISO/IEC 646-IRV characters with CARRIAGE RETURN/LINE FEED as end of line mark, that specifies, after a 6-lines header and a variable number of comment lines (starting with ‘#’), source reference lines; each line containing the following information organized in fields delimited by ‘;’:

- 1st field: UCS code point or sequence, in the format (hhhh | hhhh) (<space> (hhhh | hhhh)) *
- 2nd field: DoCoMo Shift-JIS code, in the format (hhhh)
- 3rd field: KDDI Shift-JIS code, in the format (hhhh)
- 4th field: SoftBank Shift-JIS code, in the format (hhhh)

The format definition uses ‘h’ as a hexadecimal unit and <space> as the SPACE character. An ASTERISK indicates zero, one, or more iteration of the preceding pattern.

[Click on this highlighted text to access the reference file.](#)

NOTE 1 – The content is also available as a separate viewable file in the same file directory as this document. The file is named “EmojiSrc.txt”. The content is for reference only and is not intended for cross mapping between vendor sets.

NOTE 2 – This content provides mappings between UCS code points and sequences on one hand and Shift-JIS codes for cell phone carrier symbols on the other hand. Each mapping is symmetric (“round trip”), for equivalent UCS and carrier symbols or sequences. It does not include best-fit (“fallback”) mappings to similar but not equivalent symbols in either mapping direction.

23 Source references for CJK Ideographs

23.1 List of source references

A CJK Ideograph is always referenced by at least one source reference. These source references are provided in a machine-readable format that is accessible as links to this document. The content pointed by these links is also normative.

NOTE 1 – The referenced files are only available to users who obtain their copy of the standard in a machine-readable format. However, the file format makes them printable.

The source reference information establishes the character identity for CJK Ideographs. A source reference is established by associating a CJK Ideograph code point with one or several values in the source standards listed below. Such a source standard originates from the following categories:

- Hanzi G sources,
- Hanzi H sources,
- Hanzi M sources,

ISO/IEC 10646:2014 (E)

- Hanzi T sources,
- Kanji J sources,
- Hanja K sources,
- Hanja KP sources,
- ChuNom V sources, and
- Unicode U sources

For a given code point, only one source reference can be created for each of the source standard category (G, H, M, T, J, K, KP, V, and U). In order to provide a comprehensive coverage for a source standard category, when a source standard is referenced, all its unique associations with existing CJK Ideographs are documented.

The following list identifies all sources referenced by the CJK Ideographs in both the BMP and the SIP.

NOTE 2 – Even if there is a new version of the source publication, the existing source reference information in the data files will not be updated. The updated source may only identify characters not previously covered by the older version.

The Hanzi G sources are

G0	GB2312-80
G1	GB12345-90
G3	GB7589-87 unsimplified forms
G5	GB7590-87 unsimplified forms
G7	General Purpose Hanzi List for Modern Chinese Language, and General List of Simplified Hanzi
GS	Singapore Characters
G8	GB8565-88
G9	GB18030-2000
GE	GB16500-95
GH	GB15564-1995 Code of Chinese Ideogram set for teltext broadcasting Hong Kong subset
GK	GB12052-89 Korean Character Coded Character Set for Information Interchange
G4K	Siku Quanshu (四庫全書)
GBK	Chinese Encyclopedia (中國大百科全書)
GCH	Ci Hai (辭海)
GCY	Ci Yuan (辭源)
GCCY	Chinese Academy of Surveying and Mapping Ideographs (中国测绘科学研究院用字)
GDZ	Geographic Publishing House Ideographs (地质出版社用字)
GFZ	Founder Press System (方正排版系统)
GGH	Gudai Hanyu Cidian (古代汉语词典)
GHC	Hanyu Dacidian (漢語大詞典)
GHZ	Hanyu Dazidian ideographs (漢語大字典)
GIDC	ID system of the Ministry of Public Security of China, 2009
GJZ	Commercial Press Ideographs (商务印书馆用字)
GKX	Kangxi Dictionary ideographs (康熙字典) 9 th edition (1958) including the addendum (康熙字典)補遺
GRM	People's Daily Ideographs (人民日报用字)
GXC	Xiandai Hanyu Cidian (现代汉语词典)
GXH	Xinhua Zidian (新华字典)
GWZ	Hanyu Dacidian Publishing House Ideographs (漢語大詞典出版社用字)
GZFY	Hanyu Fangyan Dacidian (汉语方言大辞典)
GZH	Zhonghua Zihai (中华字海)
GZJW	Yinzhou Jinwen Jicheng Yinde (殷周金文集成引得)

Note 3 – The graphic symbol shown on the code charts for a character referenced by a Kangxi Dictionary (GKX) are in modern Chinese style which may differ slightly from the corresponding graphic symbol used in the dictionary.

The Hanzi H sources are

H	Hong Kong Supplementary Character Set – 2008
HB0	Big-5: Computer Chinese Glyph and Character Code Mapping Table, Technical Report C-26, 電腦用中文字型與字碼對照表, 技術通報 C-26, 1984, Symbols
HB1	Big-5, Level 1
HB2	Big-5, Level 2

The Hanzi M source is

MAC	Macao Information System Character Set (澳門資訊系統字集)
-----	---

The Hanzi T sources are

T1	TCA-CNS 11643-1992 1st plane
T2	TCA-CNS 11643-1992 2nd plane
T3	TCA-CNS 11643-1992 3rd plane with some additional characters
T4	TCA-CNS 11643-1992 4th plane
T5	TCA-CNS 11643-1992 5th plane
T6	TCA-CNS 11643-1992 6th plane
T7	TCA-CNS 11643-1992 7th plane
TB	TCA-CNS 11643-2007 11th plane
TC	TCA-CNS 11643-2007 12th plane
TD	TCA-CNS 11643-2007 13th plane
TE	TCA-CNS 11643-2007 14th plane
TF	TCA-CNS 11643-2007 15th plane

The Kanji J sources are

J0	JIS X 0208-1990
J1	JIS X 0212-1990
J3	JIS X 0213:2000 level-3
J3A	JIS X 0213:2004 level-3
J4	JIS X 0213:2000 level-4
JA	Unified Japanese IT Vendors Contemporary Ideographs, 1993
JH	Hanyo-Denshi Program (汎用電子情報交換環境整備プログラム), 2002-2009
JK	Japanese KOKUJI Collection
JARIB	Association of Radio Industries and Businesses (ARIB) ARIB STD-B24 Version 5.1, March 14 2007

The Hanja K sources are

K0	KS C 5601-1987 (now known as KS X1001:2004)
K1	KS C 5657-1991 (now known as KS X1002:2001)
K2	PKS C 5700-1 1994 (Reedited and standardized as KS X1027-1:2011)
K3	PKS C 5700-2 1994 (Reedited and standardized as KS X1027-2:2011)
K4	PKS 5700-3:1998 (Reedited and standardized as KS X1027-3:2011)
K5	Korean IRG Hanja Character Set 5th Edition: 2001 (Reedited and standardized as KS X1027-4:2011)

The Hanja KP sources are

KP0	KPS 9566-97
KP1	KPS 10721:2000 and KPS 10721:2003

The ChuNom V sources are

V0	TCVN 5773:1993
V1	TCVN 6056:1995

ISO/IEC 10646:2014 (E)

V2	VHN 01:1998
V3	VHN 02: 1998
V4	Dictionary on Nom 2006, Dictionary on Nom of Tay ethnic 2006, Lookup Table for Nom in the South 1994

The Unicode U source is

UTC The Unicode Technical Report #45, U-source Ideographs, September 2012

23.2 Source references file for CJK Ideographs

The content linked to is a plain text file, using ISO/IEC 646-IRV characters with CARRIAGE RETURN/ LINE FEED as end of line mark that specifies the sources references data for all CJK Ideographs. The file also contains information concerning Radical-Stroke index, the corresponding CJK unified ideograph code point value for each CJK compatibility ideograph entry, and information about the IICORE collection (see A.4.1) for characters belonging to that collection. Each line in the text file contains the following information organized in three fields:

- 1st field: UCS code point in the format (U+hhhh) or (U+hhhhh)
- 2nd field: Tag indicating the type of information in the third field (kIRG_GSource, kIRG_HSource, kIRG_MSource, kIRG_TSource, kIRG_JSource, kIRG_KSource, kIRG_KPSource, kIRG_VSource, kIRG_USource, kIICore, kCompatibilityVariant, kRSUnicode)
- 3rd field: Information corresponding to the tag value specified by the second field. The Table 5 provides the format details.

The format definition for the first field uses 'h' as a hexadecimal unit. The three fields are delimited by a LINE TABULATION control character (000B). Comment lines, starting with a NUMBER SIGN ('#') are informational only. Comment lines and blank lines in the text file should be ignored by any automatic process which parses the data file to extract the source reference information.

Table 5: Format details of the tags used in the source reference file for CJK ideographs

Tag Value	Tag description	Third field format
kIRG_GSource	Hanzi source G	(G0-hhhh), (G1-hhhh), (G3-hhhh), (G5-hhhh), (G7-hhhh), (GS-hhhh), (G8-hhhh), (G9-hhhh), (GE-hhhh), (GH-hhhh), (GK-hhhh), (G4K), (GBK), (GBK-dddd.dd), (GCH), (GCH-dddd.dd), (GCY), (GCY-dddd.dd), (GCYY-ddddd), (GDZ-dddd.dd), (GFZ), (GFZ-ddddd), (GGH-ddddd.dd), (GHC), (GHC-dddd.dd), (GHZ-ddddd.dd), (GIDC-ddd), (GJZ-ddddd), (GKX-dddd.dd), (GRM-dddd.dd), (GXC-dddd.dd), (GXH-dddd.dd), (GWZ-dddd.dd), (GZFY-ddddd), (GZH-dddd.dd), or (GZJW-ddddd)
kIRG_HSource	Hanzi source H	(H-hhhh), (HB0-hhhh), (HB1-hhhh), or (HB2-hhhh)
kIRG_MSource	Hanzi source M	(MAC-ddddd)
kIRG_TSource	Hanzi source T	T1-hhhh), (T2-hhhh), (T3-hhhh), (T4-hhhh), (T5-hhhh), (T6-hhhh), (T7-hhhh), (TB-hhhh), (TC-hhhh), (TD-hhhh), (TE-hhhh), or (TF-hhhh)
kIRG_JSource	Kanji source J	(J0-hhhh), (J1-hhhh), (J3-hhhh), (J3A-hhhh), (J4-hhhh), (JA-hhhh), (JH-xxxxxx), (JH-xxxxxxS), (JK-ddddd), or (JARIB-hhhh)
kIRG_KSource	Hanja source K	(K0-hhhh), (K1-hhhh), (K2-hhhh), (K3-hhhh), (K4-hhhh), or (K5-hhhh)

Tag Value	Tag description	Third field format
kIRG_KPSource	Hanja KP source	(KP0-hhhh) or (KP1-hhhh)
kIRG_VSource	ChuNom V source	(V0-hhhh), (V1-hhhh), (V2-hhhh), (V3-hhhh), or (V4-hhhh)
kIRG_USource	Unicode U source	(UTC-ddddd) or (UCI-ddddd)
kIICore	IICORE info	([ABC]{1}[GTJHKMP]{1,7})
kCompatibility Variant	Compatibility info	(U+hhhh), or (U+hhhhh)
kRSUnicode	Radical Stroke index	((d{1,3}' .d{1,2})) (<space> (d{1,3}' .d{1,2}))*

The format definition uses 'd' as a decimal unit, 'h' as a hexadecimal unit, 'x' as an alphanumerical unit (0 to 9 and A to Z), and <space> as the SPACE character. Uppercase characters, digits and all other symbols between parentheses appear as shown. An ASTERISK indicates zero, one, or more iteration of the preceeding pattern.

The IICORE value ([ABC]{1}[GTJHKMP]) is a field starting with one of the letters A, B, or C indicating a decreasing order of priority, followed by one or several letters (G,T,J,H,K,M,P) indicating usage from the Hanzi G source, Hanzi T source, Kanji J source, Hanzi H source, Hanja K source, Hanzi M source, and Hanja KP source respectively. The presence of an IICORE tag for a given CJK ideograph indicates that the character is part of the IICORE collection.

The Compatibility information is a field specified for all CJK compatibility ideographs that provides the code point of the corresponding CJK unified ideograph.

The Radical-Stroke index ((d{1,3}' .d{1,2})) (space (d{1,3}' .d{1,2}))* is an informative field contains one or more of the following space separated construct: a radical index (one to three digits), optionally followed by an apostrophe for simplified radicals, followed by a full stop, and ending by one or two digits for the stroke count. Only the first Radical-Stroke is displayed in the code charts.

NOTE 1 – Concerning JIS X 0213:2000 and 2004 sources, level-4 references correspond to the second plane; other level references correspond to the first plane.

NOTE 2 – The original source references in the Hanja K4 and K5 sources are described using a single decimal index without section or position values. For better consistency with the other sources, those indexes have been converted into hexadecimal values in the source reference file. Unlike K0-K3 indexes, K4 and K5 indexes do not decompose in section, position values.

NOTE 3 – Characters using the reference type UCI-ddddd (U-source) have no identified source reference. The UCI value is simply a place holder.

NOTE 4 – The UCS code points are using the U+ prefix for UCS short identifiers in the format included in the source reference file to be identical to the similar file included in the Unicode Standard.

The following examples shows file entries for the CJK ideographs 3687, 4E00, 4E07, and F928. The three first definitions correspond to CJK unified ideographs, with 4E00 and 4E07 part of the IICORE collection, while the fourth definition corresponds to a CJK compatibility ideograph.

EXAMPLES

U+3687	kIRG_GSource	G3-3A36
U+3687	kIRG_KPSource	KP1-3C87
U+3687	kIRG_KSource	K3-2339
U+3687	kIRG_TSource	T4-2861
U+3687	kRSUnicode	35.6 66.6
U+4E00	kIRG_GSource	G0-523B
U+4E00	kIRG_HSource	HB1-A440
U+4E00	kIRG_JSource	J0-306C

ISO/IEC 10646:2014 (E)

U+4E00	kIRG_KPSource	KP0-FCD6
U+4E00	kIRG_KSource	K0-6C69
U+4E00	kIRG_TSource	T1-4421
U+4E00	kIRG_VSource	V1-4A21
U+4E00	kRSUnicode	1.0
U+4E00	kIICore	AGTJHKMP
U+4E07	kIRG_GSource	G0-4D72
U+4E07	kIRG_HSource	HB2-C945
U+4E07	kIRG_JSource	J0-4B7C
U+4E07	kIRG_KPSource	KP0-DAB9
U+4E07	kIRG_KSource	K0-5832
U+4E07	kIRG_TSource	T2-2126
U+4E07	kIRG_VSource	V1-4A24
U+4E07	kRSUnicode	1.2
U+4E07	kIICore	AGJKP
U+F928	kIRG_JSource	J3-742E
U+F928	kIRG_KSource	K0-5227
U+F928	kRSUnicode	53.9
U+F928	kCompatibilityVariant	U+5ECA

[Click on this highlighted text to access the reference file.](#)

NOTE – The content is also available as a separate viewable file in the same file directory as this document. The file is named: "CJKSrc.txt".

23.3 Source reference presentation for CJK Unified Ideographs

23.3.1 General

Unlike many other character repertoires which only show one graphic symbol per character, the code charts for CJK Unified Ideographs show one graphic symbol per source reference with additional information related to their identity such as radical, stroke count, and numeric value of the various source references.

NOTE – The presentation of the twelve CJK Unified Ideographs included in the CJK COMPATIBILITY block follows a different model described in 23.4.

The graphic representation for the radical is shown immediately below the code point, along with the radical number and the stroke count. That stroke count does not include the radical itself.

The code chart for the CJK UNIFIED IDEOGRAPHS block (4E00-9FFF) uses a fixed column format (i.e. source references from a given source always appear in the same column) while the code charts for the other CJK Unified blocks show graphic symbols per the following order of appearance: G, T, J, K, KP, V, H, M, and U.

23.3.2 Source reference presentation for CJK UNIFIED IDEOGRAPHS block

For the presentation of the CJK UNIFIED IDEOGRAPH block, the graphic representations for the Hanzi G, H, and T sources, the Kanji J source, the Hanja K source, and the ChuNom V source are shown in that order when present. Unicode U sources are shown in the second column (instead of the Hanzi H source). No characters in this block have both a H and U source.

The figure 2 shows an example for characters 4E00-4E09 and 4E14-4E1D.

HEX	C		J	K	V		HEX	C		J	K	V	
4E00 — 1.0	一	一	一	一	一	一	4E14 — 1.4	且	且	且	且	且	且
	G0-523B	HB1-A440	T1-4421	J0-306C	K0-6C69	V1-4A21		G0-4752	HB1-A542	T1-4562	J0-336E	K0-7326	V1-4A2D
4E01 — 1.1	丁	丁	丁	丁	丁	丁	4E15 — 1.4	丕	丕	丕	丕	丕	丕
	G0-3621	HB1-A442	T1-4423	J0-437A	K0-6F4B	V1-4A22		G0-5827	HB1-A541	T1-4561	J0-5023	K0-5D60	V1-4A2E
4E02 — 1.1	𠂇		𠂇	𠂇			4E16 — 1.4	世	世	世	世	世	世
	G5-3021		T4-2126	J1-3021				G0-4A40	HB1-A540	T1-4560	J0-4024	K0-6126	V1-4A2F
4E03 — 1.1	七	七	七	七	七	七	4E17 — 1.3	卅		卅	卅		
	G0-465F	HB1-A443	T1-4424	J0-3C37	K0-7652	V1-4A23		GE-2124		T4-2155	J0-5242		
4E04 — 1.1	丄	丄	丄	丄			4E18 — 1.4	丘	丘	丘	丘	丘	丘
	GE-2121	H-9EB3	T3-2126	J1-3022				G0-4770	HB1-A543	T1-4563	J0-3556	K0-4E78	V1-4A30
4E05 — 1.1	𠂇		𠂇	𠂇			4E19 — 1.4	丙	丙	丙	丙	丙	丙
	GE-2122		T3-2125	J1-3023				G0-317B	HB1-A4FE	T1-455F	J0-4A3A	K0-5C30	V1-4A31
4E06 — 1.1	𠂇				𠂇		4E1A — 1.4	业	业				
	GK-6837				K2-2121			G0-5235	H-9EB2				
4E07 — 1.2	万	万	万	万	万	万	4E1B — 1.4	丛					
	G0-4D72	HB2-C945	T2-2126	J0-4B7C	K0-5832	V1-4A24		G0-3454					
4E08 — 1.2	丈	丈	丈	丈	丈	丈	4E1C — 1.4	东	东				
	G0-5549	HB1-A456	T1-4437	J0-3E66	K0-6D5B	V1-4A25		G0-362B	H-9DD6				
4E09 — 1.2	三	三	三	三	三	三	4E1D — 1.4	丝					
	G0-487D	HB1-A454	T1-4435	J0-3B30	K0-5F32	V1-4A26		G0-4B3F					

Figure 2 – Chart presentation for CJK UNIFIED IDEOGRAPHS

23.3.3 Source reference presentation for CJK UNIFIED IDEOGRAPHS EXTENSION A

For the presentation of the CJK UNIFIED IDEOGRAPH EXTENSION A block, up to three sources per characters are represented in a single row. If more than three sources exist, an additional row is used.

The figure 3 shows an example for characters 41C9-41CC, 41DB-41DD, and 41EE-41F0.

41C9 立 117.5	𪛗	𪛗		41DB 竹 118.4	𪛗	𪛗	𪛗	41EE 竹 118.6	𪛗	𪛗	𪛗
	GHZ-42707.25	T3-3322			GKX-0879.12	T3-3329	V2-7F4B		G5-6334	T4-3975	JA-254D
41CA 立 117.5	𪛗	𪛗			𪛗				𪛗		
	JA-2549	H-8E55			H-8EFE				V2-7F50		
41CB 立 117.6	𪛗	𪛗	𪛗	41DC 竹 118.4	𪛗	𪛗	𪛗	41EF 竹 118.6	𪛗	𪛗	𪛗
	GKX-0871.06	T5-3446	JA-254A		G3-634F	T4-2E73	K3-2E2D		G5-632E	T3-3D73	H-8E59
41CC 立 117.7	𪛗	𪛗		41DD 竹 118.4	𪛗	𪛗		41F0 竹 118.6	𪛗	𪛗	
	GKX-0871.17	T3-3D6F			G3-634A	T4-2E72			G3-6379	T4-396F	

Figure 3 – Chart presentation for CJK UNIFIED IDEOGRAPHS EXTENSION A

23.3.4 Source reference presentation for CJK UNIFIED IDEOGRAPHS EXTENSION B

For the presentation of the CJK UNIFIED IDEOGRAPH EXTENSION B block, the first graphic symbol shows the glyph used for the first and second edition of this International Standard (2003 and 2011 re-

spectively) referenced by a 'UCS2003' notation. Up to two sources per characters are represented in a single row. If more than two sources exist, an additional row is used.

The figure 4 shows an example for characters 200E8-200EB, 200FC-200FF, and 2010E-20110.

200E8 乙 5.5 𠂇 𠂇 UCS2003 GHZ-10053.11	200FC 乙 5.10 𠂇 𠂇 𠂇 UCS2003 GKX-0084.29 T6-3C7E	2010E J 6.1 𠂇 𠂇 𠂇 UCS2003 GKX-0085.13 T4-2128
200E9 乙 5.6 𠂇 𠂇 UCS2003 V0-305E	200FD 乙 5.10 𠂇 𠂇 𠂇 UCS2003 GKX-0084.30 T5-3073	𠂇 H-8853
200EA 乙 5.6 𠂇 𠂇 UCS2003 TF-2275	200FE 乙 5.10 𠂇 𠂇 𠂇 UCS2003 GKX-0084.32 T6-3C7D	2010F J 6.1 𠂇 𠂇 𠂇 UCS2003 GKX-0085.14 T5-2128
200EB 田 102.2 𠂇 𠂇 UCS2003 GHZ-10053.14	200FF 乙 5.10 𠂇 𠂇 UCS2003 GHZ-10057.03	20110 J 6.1 𠂇 𠂇 UCS2003 K4-0001

Figure 4 – Chart presentation for CJK UNIFIED IDEOGRAPHS EXTENSION B

23.3.5 Source reference presentation for CJK UNIFIED IDEOGRAPHS EXTENSION C, D and E

For the presentation of the CJK UNIFIED IDEOGRAPH EXTENSION C, D and E blocks, up to two sources per characters are represented in a single row. If more than two sources exist, an additional row is used.

The figure 5 shows an example for characters 2A7A0-2A7A2, 2A7B4-2A7B6, 2A7C8-2A7CA, and 2A7DC-2A7DE.

2A7A0 彳 15.7 𠂇 V4-416A	2A7B4 儿 16.9 𠂇 V4-4175	2A7C8 刀 18.8 𠂇 𠂇 GZJW-00095 TC-4456	2A7DC 力 19.7 𠂇 TC-3E36
2A7A1 彳 15.7 𠂇 TC-3770	2A7B5 儿 16.11 𠂇 JK-65041	2A7C9 刀 18.8 𠂇 TC-445D	2A7DD 力 19.8 𠂇 GXC-3001.22
2A7A2 彳 15.8 𠂇 TC-443B	2A7B6 匚 17.3 𠂇 UTC-00094	2A7CA 刀 18.9 𠂇 TC-5261	2A7DE 力 19.8 𠂇 V4-4244

Figure 5 – Chart presentation for CJK UNIFIED IDEOGRAPHS EXTENSION C, D and E

23.4 Source references presentation for CJK Compatibility Ideographs

Similarly to CJK Unified Ideographs, the code charts for CJK Compatibility Ideographs show one graphic symbol per source reference with additional information related to their identity such as radical, stroke count, and numeric value of the various source references.

The graphic representation for the radical is also shown immediately below the code point, along with the radical number and the stroke count. That stroke count does not include the radical itself.

Additional information, corresponding to informative items described in 31.3, such as decomposition mapping, may also appear in the presentation.

NOTE – The presentation of the twelve CJK Unified Ideographs included in the CJK COMPATIBILITY IDEOGRAPHS block follows this model. It does not include decomposition mapping.

The figure 6 shows an example for characters FA0C-FA0E, FA19-FA1D, and FA28-FA2C.

<p>Duplicate characters from Big 5</p> <p>FA0C 兀 兀 儿 10.1 HB2-C94A UTC-00915 ≡ 5140 兀</p> <p>FA0D 殼 殼 口 30.10 HB2-DDFC UTC-00916 ≡ 55C0 殼</p> <p>The IBM 32 compatibility ideographs</p> <p><i>This section contains twelve characters that are CJK unified ideographs. Each CJK unified ideograph is annotated as such.</i></p> <p>FA0E 雥 雥 又 29.11 UTC-00843 • a CJK unified ideograph</p>	<p>FA19 示 113.5 神 神 J3-793C UTC-00923 ≡ 795E 神</p> <p>FA1A 示 113.6 祥 祥 J3-793D UTC-00924 ≡ 7965 祥</p> <p>FA1B 示 113.9 福 福 J3-7941 UTC-00925 ≡ 798F 福</p> <p>FA1C 青 174.5 靖 靖 UTC-00926 ≡ 9756 靖</p> <p>FA1D 米 119.8 精 精 UTC-00927 ≡ 7CBE 精</p>	<p>FA28 金 167.8 鋹 鋹 TF-584C UTC-00853 • a CJK unified ideograph</p> <p>FA29 阜 170.10 隄 隄 UTC-00854 • a CJK unified ideograph</p> <p>FA2A 食 184.4 飯 飯 UTC-00933 ≡ 98EF 飯</p> <p>FA2B 食 184.5 飼 飼 UTC-00934 ≡ 98FC 飼</p> <p>FA2C 食 184.8 館 館 UTC-00935 ≡ 9928 館</p>
--	--	---

Figure 6 – Chart presentation for CJK COMPATIBILITY IDEOGRAPHS

24 Character names and annotations

24.1 Entity names

This standard specifies names for the following entity types

- Characters, as character names and character name aliases
- named UCS sequences identifiers (see Clause 25)
- blocks (see Clause 14 and A.2)
- collections (see A.1)

The names given by this standard to these entities shall follow the rules for name formation and name uniqueness specified in Clause 24. This specification applies to the entity names in the English language version of this standard.

NOTE 1 – In a version of such a standard in another language a) these rules may be amended to permit names to be generated using words and syntax that are considered appropriate within that language; b) the entity names from this version of the standard may be replaced by equivalent unique names constructed according to the rules amended as in a) above.

NOTE 2 – Additional guidelines for constructing entity names are given in Annex L for information.

24.2 Name formation

An entity name shall consist only of the following characters

- LATIN CAPITAL LETTER A through LATIN CAPITAL LETTER Z,
- DIGIT ZERO through DIGIT NINE,
- SPACE,
- HYPHEN-MINUS, and
- FULL STOP if the entity being named is a collection

The first character in an entity name shall be a Latin capital letter. The last character in an entity name shall be either a Latin capital letter or a Digit.

An entity name shall not contain two or more consecutive SPACE characters or consecutive HYPHEN-MINUS characters. Furthermore, except for collection names, an entity name shall not contain a digit

(DIGIT ZERO through DIGIT NINE) preceded by a SPACE character. A collection name shall not contain two or more consecutive FULL STOP characters.

A sequence of a SPACE followed by a HYPHEN-MINUS or a sequence of a HYPHEN-MINUS followed by a SPACE may appear only in character names or named UCS sequence identifiers.

EXAMPLE 1 Each of the following two character names contains a consecutive SPACE and HYPHEN-MINUS:

TIBETAN LETTER -A

TIBETAN MARK BKA- SHOG YIG MGO

FULL STOP may appear only in between two alpha-numeric characters (LATIN CAPITAL LETTER A through LATIN CAPITAL LETTER Z, DIGIT ZERO through DIGIT NINE) in a collection name.

EXAMPLE 2 The following collection name contains FULL STOP in between two Digits, DIGIT FOUR and DIGIT ONE:

UNICODE 4.1

EXAMPLE 3 The following collection name contains FULL STOP in between one Latin letter, LATIN CAPITAL LETTER D, and a Digit, DIGIT SEVEN:

BMP-AMD.7

24.3 Single name

Each entity named in this standard shall be given only one name. However, one or more character name aliases may also be associated with a character.

NOTE – This does not preclude the informative use of name aliases or acronyms for the sake of clarity. However, the normative entity name will be unique. These informative aliases should not be confused with character name aliases which share the same name space as character names and are normative.

24.4 Name immutability

Some entity names cannot be changed by future versions of this standard. This applies to character name and character name aliases. See Clause 7.

24.5 Name uniqueness

Each entity name shall also be unique within an appropriate name space, as specified here.

24.5.1 Block names

Block names constitute a name space. Each block name shall be unique and distinct from all other block names specified in the standard.

24.5.2 Collection names

Collection names constitute a name space. Each collection name shall be unique and distinct from all other collection names specified in the standard.

24.5.3 Character names, character name aliases, and named UCS sequence identifiers

Character names, character name aliases and named UCS sequence identifiers, taken together, constitute a name space. Each character name, character name aliases, or named UCS sequence identifier shall be unique and distinct from all other character names, character name aliases, or named UCS sequence identifiers.

24.5.4 Determining uniqueness

For block names and collection names, two names shall be considered unique and distinct if they are different even when SPACE and medial HYPHEN-MINUS characters are ignored in comparison of the names. A medial HYPHEN-MINUS is a HYPHEN-MINUS character that occurs immediately after a character other than SPACE and immediately before a character other than SPACE.

EXAMPLE 1 The following hypothetical block names would be unique and distinct:

LATIN-A

LATIN-B

EXAMPLE 2 The following hypothetical block names would not be unique and distinct:

LATIN-A
LATIN A
LATINA

For character names, character name aliases, and named UCS sequence identifiers, two names shall be considered unique and distinct if they are different even when SPACE and medial HYPHEN-MINUS characters are ignored and even when the words "LETTER", "CHARACTER", and "DIGIT" are ignored in comparison of the names.

EXAMPLE 3 The following hypothetical character names would not be unique and distinct:

MANICHAEAN CHARACTER A
MANICHAEAN LETTER A

EXAMPLE 4: The following two actual character names are unique and distinct, because they differ by a HYPHEN-MINUS that is not a medial HYPHEN-MINUS:

TIBETAN LETTER A
TIBETAN LETTER -A

The following two character names shall be considered unique and distinct:

HANGUL JUNGSEONG OE
HANGUL JUNGSEONG O-E

NOTE – These two character names are explicitly handled as an exception, because they were defined in an earlier version of this International Standard before the introduction of the name uniqueness requirement. This pair is, has been, and will be the only exception to the uniqueness rule in this International Standard.

24.6 Character names for CJK Ideographs

For CJK Ideographs the names are algorithmically constructed by appending their coded representation in hexadecimal notation to "CJK UNIFIED IDEOGRAPH-" for CJK Unified Ideographs and "CJK COMPATIBILITY IDEOGRAPH-" for CJK Compatibility Ideographs.

For CJK Ideographs within the BMP, the coded representation is their two-octet value expressed as four hexadecimal digits. For example, the first CJK Ideograph character in the BMP has the name "CJK UNIFIED IDEOGRAPH-3400".

For CJK Ideographs within the SIP, the coded representation is their five hexadecimal digit value. For example, the first CJK Ideograph character in the SIP has the name "CJK UNIFIED IDEOGRAPH-20000".

24.7 Character names for Hangul syllables

Names for the Hangul syllable characters in code points AC00 - D7A3 are derived from their code point values by the numerical procedure described below. Lists of names for these characters are not provided in the code charts.

- Obtain the code point of the Hangul syllable character. It is of the form $h_1h_2h_3h_4$ where h_1 , h_2 , h_3 , and h_4 are hexadecimal digits representing the number $h_1h_2h_3h_4$ lying within the range AC00 to D7A3.
- Derive the decimal numbers d_1 , d_2 , d_3 , d_4 that are numerically equal to the hexadecimal digits h_1 , h_2 , h_3 , h_4 respectively.
- Calculate the character index C from the formula

$$C = 4096 \times (d_1 - 10) + 256 \times (d_2 - 12) + 16 \times d_3 + d_4$$
- Calculate the syllable component indices I , P , F from the following formulae

$$I = C / 588 \quad (\text{Note: } 0 \leq I \leq 18)$$

$$P = (C \% 588) / 28 \quad (\text{Note: } 0 \leq P \leq 20)$$

$$F = C \% 28 \quad (\text{Note: } 0 \leq F \leq 27)$$
 where "/" indicates integer division (i.e. x / y is the integer quotient of the division), and "%" indicates the modulo operation (i.e. $x \% y$ is the remainder after the integer division x / y).
- Obtain the Latin character strings that correspond to the three indices I , P , F from columns 2, 3, and 4 respectively of table 1 below (for $I = 11$ and for $F = 0$ the corresponding strings are null). Concatenate these three strings in left-to-right order to make a single string, the syllable-name.

- f) The character name for the character code point $h_1h_2h_3h_4$ is then HANGUL SYLLABLE $s-n$ where “ $s-n$ ” indicates the syllable-name string derived in step e.

EXAMPLE

For the character with code point D4DE:

$d_1 = 13, d_2 = 4, d_3 = 13, d_4 = 14$.

$C = 10462$

$I = 17, P = 16, F = 18$.

The corresponding Latin character strings are P, WI, BS. The syllable-name is PWIBS, and the character name is HANGUL SYLLABLE PWIBS.

For each Hangul syllable character, a short annotation is also defined. This annotation consists of an alternative transliteration of the Hangul syllable into Latin characters. They are also derived from their code point values by a similar numerical procedure described below.

- g) Carry out steps a to d as described above.
- h) Obtain the Latin character strings that correspond to the three indices I, P, F from columns 5, 6, and 7 respectively of Table 6 below (for $I = 11$ and for $F = 0$ the corresponding strings are null). Concatenate these three strings in left-to-right order to make a single string.

EXAMPLE

For the character with code point D4DE:

$d_1 = 13, d_2 = 4, d_3 = 13, d_4 = 14$.

$C = 10462$

$I = 17, P = 16, F = 18$.

The corresponding Latin character strings are ph, wi, ps; and the annotation is phwips.

The Annex R provides the names and annotation of Hangul syllables through a linked file.

Table 6: Elements of Hangul syllable names and annotation

Index number	Syllable name elements			Annotation elements		
	I string	P string	F string	I string	P string	F string
0	G	A		k	a	
1	GG	AE	G	kk	ae	k
2	N	YA	GG	n	ya	kk
3	D	YAE	GS	t	yae	ks
4	DD	EO	N	tt	eo	n
5	R	E	NJ	r	e	nc
6	M	YEO	NH	m	yeo	nh
7	B	YE	D	p	ye	t
8	BB	O	L	pp	o	l
9	S	WA	LG	s	wa	lk
10	SS	WAE	LM	ss	wae	lm
11		OE	LB		oe	lp
12	J	YO	LS	c	yo	ls
13	JJ	U	LT	cc	u	lth
14	C	WEO	LP	ch	weo	lph
15	K	WE	LH	kh	we	lh
16	T	WI	M	th	wi	m
17	P	YU	B	ph	yu	p
18	H	EU	BS	h	eu	ps
19		YI	S		yi	s
20		I	SS		i	ss
21			NG			ng
22			J			c
23			C			ch

24			K			kh
25			T			th
26			P			ph
27			H			h

NOTE 1 – The I and F strings in Syllable name elements of Table 5 correspond to the Hangul Jamo short names shown in annotations in the code chart after the names of the Hangul Jamo characters in the ranges 1100 to 1112 (except 110B) and 11A8 to 11C2. The short names are transliterations of these Hangul characters.

NOTE 2 – The I and F strings in Syllable name elements are based on Method I of ISO/TR 11941:1996 Information and documentation – Transliteration of Korean script into Latin characters. The I and F strings in Annotation elements of the same Table 5 are based on Method II of ISO/TR 11941. P strings in Syllable name elements and in Annotation elements are based on ISO/TR 11941. ISO/TR 11941 is different from Revised Romanization of Korean script released on 4 July 2000 by the Ministry of Culture and Tourism, Republic of Korea.

25 Named UCS Sequence Identifiers

A Named UCS Sequence Identifier (NUSI) is a USI associated to a name following the same construction rules as for character names. These rules are given in 24.

NOTE 1 – The purpose of these named USIs is to specify sequences of characters that may be treated as single units, either in particular types of processing, in reference by standards, in listing of repertoires (such as for fonts or keyboards).

The USI value corresponding to each NUSI is written using the coded representation determined by the Normalization Form C (NFC) (see Clause 21). Each named UCS sequence has a unique code representation. All allowed named UCS sequence identifiers for use with this International Standard are specified in the content linked below. All other such named sequences are undefined.

The content linked to is a plain text file, using ISO/IEC 646-IRV characters with CARRIAGE RETURN/LINE FEED as end of line mark that specifies Named UCS Sequence Identifiers. Each line in the text file contains the following information organized in two fields:

- 1st field: Name of the NUSI (following rules given in 24)
- 2nd field: The USI associated with that Name (using a modified syntac for the USI, omitting commas and angle brackets)

The two fields are delimited by a SEMICOLON (;) followed optionally by zero or more SPACE characters. Comment lines, starting with a NUMBER SIGN (#) are informational only. Comment lines and blank lines in the text file should be ignored by any automatic process which parses the data file to extract the normative list of NUSIs.

[Click on this highlighted text to access the reference file.](#)

NOTE 2 – The content is also available as a separate viewable file in the same file directory as this document. The file is named “NUSI.txt”.


NOTE 3 – All the allowed Named UCS Sequence Identifiers for use with this International Standard are also listed in the Unicode character database: <http://www.unicode.org/Public/UNIDATA/NamedSequences.txt>.

26 Structure of the Basic Multilingual Plane

An overview of the Basic Multilingual Plane is shown in figure 7 and a more detailed overview of Rows 00 to 33 is shown in figure 8. The Basic Multilingual Plane includes characters in general use in alphabetic, syllabic, and ideographic scripts together with various symbols and digits.

Row

00	Rows 00 to 33 (see figure 8)										
33											
34	CJK Unified Ideographs Extension A										
4D										Yijing Hexagram Symbols	
4E											
	CJK Unified Ideographs										
9F											
A0..	Yi Syllables										
A3											
A4									Yi Radicals	Lisu	
A5	Vai										
A6	Cyrillic Extended-B						Bamum				
A7	Modifier T L	Latin Extended-D									
A8	Syloti Nagri	CINF	Phags-Pa				Saurashtra			Dev Ext	
A9	Kayah Li	Rejang		Hangul Jamo Ext-A			Javanese			Myanmar EE	
AA	Cham			Myanmar Ext-A			Tai Viet			Meetei M.E.	
AB	Ethiopic Ext A	Latin Extended-E							Meetei Mayek		
AC											
	Hangul Syllables										
D7										Hangul Jamo Extended-B	
D8..	Surrogate (for use in UTF-16 only)										
DF											
E0											
	Private Use Area										
F8											
F9	CJK Compatibility Ideographs										
FA											
FB	Alphabetic Presentation Forms										
FC	Arabic Presentation Forms-A										
FD											
FE	VS	VF	CHM	CJK CF	Small Form Variants		Arabic Presentation Forms-B				
FF	Halfwidth And Fullwidth Forms										Sp.

 = permanently reserved

 = reserved for future standardization

NOTE – Vertical boundaries within rows are indicated in approximate positions only. Block names in the figure may be abbreviated due to space limitations. See A.2 for unabridged names.

Figure 7 - Overview of the Basic Multilingual Plane

Row

00	Controls	Basic Latin			Controls	Latin-1 Supplement			
01	Latin Extended-A					Latin Extended-B			
02	Latin Extended-B		IPA (Intl. Phonetic Alphabet) Extensions			Spacing Modifier Letters			
03	Combining Diacritical Marks			Greek and Coptic					
04	Cyrillic								
05	Cyrillic Supplement		Armenian			Hebrew			
06	Arabic								
07	Syriac		Arabic Sup.		Thaana		Nko		
08	Samaritan		Mandaic				Arabic Extended-A		
09	Devanagari				Bengali				
0A	Gurmukhi				Gujarati				
0B	Oriya				Tamil				
0C	Telugu				Kannada				
0D	Malayalam				Sinhala				
0E	Thai				Lao				
0F	Tibetan								
10	Myanmar					Georgian			
11	Hangul Jamo								
12	Ethiopic								
13					Ethiopic Sup.	Cherokee			
14..	Unified Canadian Aboriginal Syllabics								
16					Ogham	Runic			
17	Tagalog	Hanunoo	Buhid	Tagbanwa	Khmer				
18	Mongolian							UCAS Extended	
19	Limbu		Tai Le		New Tai Lue ^a			Khmer Symb.	
1A	Buginese	Thai Tham						Combining Diacritical M Extended	
1B	Balinese				Sundanese		Batak		
1C	Lepcha		Ol Chiki				Sund. S.	Vedic Extensions	
1D	Phonetic Extensions				Phonetic Extensions Sup.		Combining Diacritical M Sup.		
1E	Latin Extended Additional								
1F	Greek Extended								
20	General Punctuation			Super-/Subscripts		Currency Symbols		Comb. Mks. Symb.	
21	Letterlike Symbols		Number Forms			Arrows			
22	Mathematical Operators								
23	Miscellaneous Technical								
24	Control Pictures		O.C.R.	Enclosed Alphanumerics					
25	Box Drawing			Block Elements		Geometric Shapes			
26	Miscellaneous Symbols								
27	Dingbats					Misc. Math. Symbols-A		S A A	
28	Braille Patterns								
29	Supplemental Arrows-B			Miscellaneous Mathematical Symbols-B					
2A	Supplemental Mathematical Operators								
2B	Miscellaneous Symbols and Arrows								
2C	Glagolitic		Latin Ext-C		Coptic				
2D	Georgian Sup.	Tifinagh			Ethiopic Extended			Cyrillic Ext-A	
2E	Supplemental Punctuation				CJK Radicals Supplement				
2F	Kangxi Radicals							Ideog. Descr.	
30	CJK Symbols And Punctuation		Hiragana			Katakana			
31	Bopomofo	Hangul Compatibility Jamo			Kanbun	Bopomofo E.	CJK Strokes	K P E	
32	Enclosed CJK Letters And Months								
33	CJK Compatibility								

= reserved for future standardization

^a New Tai Lue is also known as Xishuang Banna Dai

NOTE – Vertical boundaries within rows are indicated in approximate positions only. Block names in the figure may be abbreviated due to space limitations. See A.2 for unabridged names.

Figure 8 - Overview of Rows 00 to 33 of the Basic Multilingual Plane

27 Structure of the Supplementary Multilingual Plane for scripts and symbols (SMP)

Because another supplementary plane is reserved for additional CJK Ideographs, the SMP (plane 01) is not used to date for encoding CJK Ideographs. Instead, the SMP is used for encoding graphic characters used in other scripts of the world that are not encoded in the BMP. Most, but not all, of the scripts encoded to date in the SMP are not in use as living scripts by modern user communities.

NOTE – The following subdivision of the SMP has been proposed:

Alphabetic scripts,
Hieroglyphic, ideographic and syllabaries,
Non CJK ideographic scripts,
Newly invented scripts,
Symbol sets

An overview of the Supplementary Multilingual Plane for scripts and symbols is shown in figure 9 and a more detailed overview of Rows 00 to 7F is shown in figure 10.

Row

00	Rows 00 to 7F (See Figure 10)		
...			
7F			
80			
...			
B0	Kana supplement		
...			
BC	Duployan	S F C	
...			
D0	Byzantine Musical Symbols		
D1	Western Musical Symbols		
D2	Ancient Greek Musical Not.		
D3	Tai Xuan Jing Symbols	Counting Rod Num	
D4			
...	Mathematical Alphanumeric Symbols		
D7			
D8	Sutton SignWriting		
...			
DA			
...			
E8	Mende		
...			
EE	Arabic Mathematical Alphabetic Symbols		
EF			
F0	Mahjong Tiles	Domino Tiles	Playing Cards
F1	Enclosed Alphanumeric Supplement		
F2	Enclosed Ideographic Supplement		
F3			
...	Miscellaneous Symbols and Pictographs		
F5			
F6	Emoticons	Ornamental Dingbats	Transport and Map Symbols
F7	Alchemical Symbols	Geometric Shapes Extended	
F8	Supplemental Arrows-C		
...			
FF			

 = reserved for future standardization

NOTE – Vertical boundaries within rows are indicated in approximate positions only. Block names in the figure may be abbreviated due to space limitations. See A.2 for unabridged names.

Figure 9 – Overview of the Supplementary Multilingual Plane for scripts and symbols

Row

00	Linear B Syllabary				Linear B Ideograms		
01	Aegean Numbers		Ancient Greek Numbers		Ancient Symbols		Phaistos Disc
02					Lycian	Carian	
03	Old Italic	Gothic	Old Permic		Ugaritic	Old Persian	
04	Deseret		Shavian		Osmanya		
05	Elbasan	Caucasian Albanian					
06							
07							
08	Cypriot Syllabary		mp Aram	Palmyrene	Nabataean		Hatran
09	Phoenician	Lydian			Meroitic Hier.	Meroitic Cursive	
0A	Kharoshthi			O South Arabian	O North Arabian	Manichaean	
0B	Avestan		Parthian	Inscr. Pahlavi	Psalter Pahlavi		
0C	Old Turkic				Hungarian		
0D							
0E					Rumi Num S.		
0F							
10	Brahmi				Kaithi		Sora Sompeng
11	Chakma		Mahajani		Sharada		Sinhala Arch N
12	Khojki				Multani	Khudawadi	
13	Grantha						
14					Tirhuta		
15					Siddham		
16	Modi				Takri		
17	Ahom						
18					Warang Citi		
19							
1A							Pau Cin Hau
...							
20							
...							
23	Cuneiform						
24	Cuneiform Numbers and Punctuation				Early Dynastic Cuneiform		
25	Early Dyn. Cun. (cont)						
...							
30	Egyptian Hieroglyphs						
...							
34							
...							
40	Anatolian Hieroglyphs						
...							
46							
...							
68	Bamum Supplement						
...							
6A			Mro			Bassa Vah	
6B	Pahawh Hmong						
...							
6F	Miao						
...							
7F							

= reserved for future standardization

NOTE – Vertical boundaries within rows are indicated in approximate positions only. Block names in the figure may be abbreviated due to space limitations. See A.2 for unabridged names.

Figure 10 – Overview of the Rows 00 to 7F of the Supplementary Multilingual Plane for scripts and symbols

28 Structure of the Supplementary Ideographic Plane (SIP)

The SIP (plane 02) is used for CJK unified ideographs (unified East Asian ideographs) that are not encoded in the BMP. The procedures for the unification and the rules for their arrangement are described in Annex S.

The SIP is also used for CJK compatibility ideographs. These ideographs are compatibility characters as specified in Clause 18.

The figure 11 shows an overview of the Supplementary Ideographic Plane.

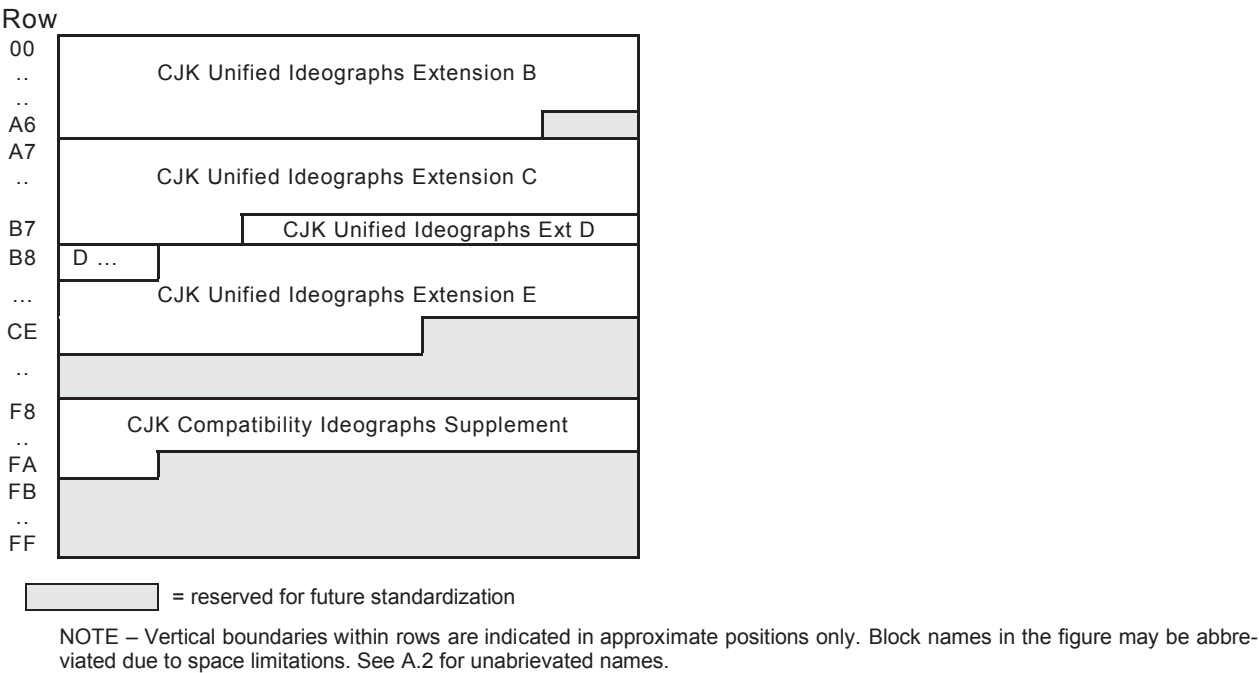


Figure 11 – Overview of the Supplementary Ideographic Plane

29 Structure of the Tertiary Ideographic Plane (TIP)

The TIP (plane 03) is used for ancient ideographic scripts that are related to but not classified as CJK unified ideographs. No characters are currently encoded in the TIP.

NOTE – The TIP may contain scripts such as Oracle Bone or Bronze in future editions of this International Standard.

30 Structure of the Supplementary Special-purpose Plane (SSP)

The SSP (plane 0E) is used for special purpose use graphic characters and format characters. The figure 12 shows an overview of the Supplementary Special-purpose Plane.

NOTE – Unassigned code points in this range should be ignored in normal processing and display.

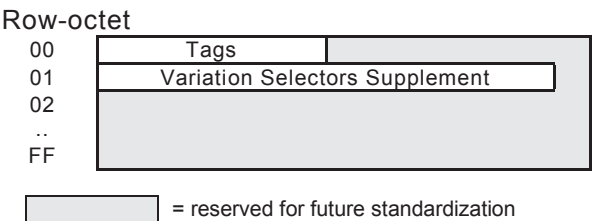


Figure 12 – Overview of the Supplementary Special-purpose Plane

31 Code charts and lists of character names

31.1 General

Detailed code charts and lists of character names for the BMP, SMP, SIP and SSP are shown on the following pages. Code charts are arranged by blocks which may span several pages.

Each code chart is followed by a corresponding character names list, except blocks for the CJK ideographs and Hangul syllables.

NOTE – A block code chart and name list may be arranged in a single page if their contents allow it.

31.2 Code chart

Code charts are presented in arrays of graphic symbols representing the characters organized in one to sixteen columns of sixteen symbols each. The lower digit of the coded representation is indicated in the left margin while the remaining upper digits are indicated in the top margin. The full coded representation for each character is also indicated under each representative graphic symbol. Code charts for CJK ideographs have different formats. See Clause 23.

NOTE – Graphic symbols corresponding to the representation of graphic characters are informative. See Clause 13.

31.3 Character names list

The character names lists contain both normative and informative information. The following information items are normative:

- Character code point,
- Associated character name,
- Character name alias (one preceded by '※').

All other information is informative and may contain:

- Graphic symbol associated with the character,
- Subheads grouping various parts of a given block. For example, the LATIN-1 SUPPLEMENT block contain “Latin-1 punctuation and symbols”, “Letters”, and “Mathematical operator”,
- Explanatory text describing context for a subhead or a whole block,
- Informative aliases, preceded by ‘=’, indicating alternate names for characters,
- Cross references, preceded by ‘→’, indicating a related character of interest,
- Information about languages, preceded by ‘•’, indicating a non exhaustive list of languages using that character. For bicameral scripts, the information is only provided for the lower case form of the character,
- Case mappings, also preceded by ‘•’, only when it cannot be derived simply from the names,
- Other information about a character, also preceded by ‘•’, describing name peculiarity, historical consideration, or any noteworthy aspect of a character,
- Decomposition mappings, preceded by ‘≡’ for canonical mappings, and by ‘≈’ for compatibility mappings,

NOTE – The decomposition mappings syntax is described in more details in section 24.1 of the Unicode Standard Version 7.0.

- Standardized variation sequences preceded by U+2053 ~ SWUNG DASH, when this character is used as a base character in such a variation sequence.

The following example describes various fragments of name lists including these informative items.

EXAMPLE

Latin-1 punctuation and symbols

Based on ISO/IEC 8859-1 (aka Latin-1) from here.

...

00B5 μ MICRO SIGN

≈ 03BC μ greek small letter mu

00B6 ¶ PILCROW SIGN

= paragraph sign

• section sign in some European usage

→ 204B ⁂ reverse pilcrow sign

→ 2761 ¶ curve stern paragraph sign ornament

...

Letters

...

00E5 å LATIN SMALL LETTER A WITH RING ABOVE

• Danish, Norwegian, Swedish, Walloon

≡ 0061 a 030A °

00E6 æ LATIN SMALL LETTER AE

= latin small ligature (1.0)

= ash (from Old English æsc)

• Danish, Norwegian, Icelandic, Faroese, Old English, French, IPA

→ 0153 œ latin small ligature oe

→ 04D5 æ cyrillic small ligature a ie

...

01C9 Ĳ LATIN SMALL LETTER LJ

→ 0459 Ѧ cyrillic small letter lje

≈ 006C I 006A j

...

2268 ⩵ LESS-THAN BUT NOT EQUAL TO

~ 2268 FE00 ⩵ with vertical stroke

...

FE18 ⎵ PRESENTATION FORM FOR VERTICAL RIGHT WHITE LENTICULAR BRACKET

⌵ PRESENTATION FORM FOR VERTICAL RIGHT WHITE LENTICULAR BRACKET

• misspelling of "BRACKET" in character name is a known defect

≈ <vertical> 3017 ⌵

31.4 Summary of standardized variation sequences

A summary of standardized variation sequences (see 16.6.2) are presented after each block in which they appear for the following categories:

- Mathematical symbols
- Mongolian characters
- Phags-pa characters
- Manichaean characters

NOTE – The summaries of standardized variation sequences corresponding to Pictographic symbols and CJK unified ideographs may be included in a future edition of this International Standard.

31.5 Pointers to code charts and lists of character names

Access to the code charts and lists of character names is provided by clicking on the appropriate highlighted text below.

- [Basic Latin to Yijing Hexagram Symbols \(0000-4DFF\)](#)
- [CJK Unified Ideographs \(4E00-9FFF\)](#)

- Yi Syllables to Alchemical Symbols (A000-1FFFF)
- CJK Unified Ideographs Extension B Part 1 of 2 (20000-25333)
- CJK Unified Ideographs Extension B Part 2 of 2 (25334-2A6FF)
- CJK Unified Ideographs Extensions C to Variation Selectors Supplement (2A700-10FFFF)

NOTE – To preserve the odd-even layout of the code charts, a page from the previous block may be inserted before the actual start of a block.

(This page left intentionally blank)

Annex A (normative) Collections of graphic characters for subsets

A.1 Collections of coded graphic characters

The collections listed below are ordered by collection number. An * in the “code points” column indicates that the collection is a fixed collection.

Collection number and name	Code points			
1 BASIC LATIN	0020-007E *	35	COMBINING DIACRITICAL MARKS FOR SYMBOLS	20D0-20FF
2 LATIN-1 SUPPLEMENT	00A0-00FF *	36	LETTERLIKE SYMBOLS	2100-214F *
3 LATIN EXTENDED-A	0100-017F *	37	NUMBER FORMS	2150-218F
4 LATIN EXTENDED-B	0180-024F *	38	ARROWS	2190-21FF *
5 IPA EXTENSIONS	0250-02AF *	39	MATHEMATICAL OPERATORS	2200-22FF *
6 SPACING MODIFIER LETTERS	02B0-02FF *	40	MISCELLANEOUS TECHNICAL	2300-23FF
7 COMBINING DIACRITICAL MARKS	0300-036F *	41	CONTROL PICTURES	2400-243F
8 BASIC GREEK	0370-03CF	42	OPTICAL CHARACTER RECOGNITION	2440-245F
9 GREEK SYMBOLS AND COPTIC	03D0-03FF	43	ENCLOSED ALPHANUMERICS	2460-24FF *
10 CYRILLIC	0400-04FF *	44	BOX DRAWING	2500-257F *
11 ARMENIAN	0530-058F	45	BLOCK ELEMENTS	2580-259F *
12 BASIC HEBREW	05D0-05EA *	46	GEOMETRIC SHAPES	25A0-25FF *
13 HEBREW EXTENDED	0590-05CF 05EB-05FF	47	MISCELLANEOUS SYMBOLS	2600-26FF *
14 BASIC ARABIC	0600-065F	48	DINGBATS	2700-27BF *
15 ARABIC EXTENDED	0660-06FF *	49	CJK SYMBOLS AND PUNCTUATION	3000-303F *
16 DEVANAGARI	0900-097F *	50	HIRAGANA	3040-309F
	200C, 200D	51	KATAKANA	30A0-30FF *
17 BENGALI	0980-09FF 200C, 200D	52	BOPOMOFO	3100-312F 31A0-31BF
18 GURMUKHI	0A00-0A7F 200C, 200D	53	HANGUL COMPATIBILITY JAMO	3130-318F
19 GUJARATI	0A80-0AFF 200C, 200D	54	CJK MISCELLANEOUS	3190-319F
20 ORIYA	0B00-0B7F 200C, 200D	55	ENCLOSED CJK LETTERS AND MONTHS	3200-32FF
21 TAMIL	0B80-0BFF 200C, 200D	56	CJK COMPATIBILITY	3300-33FF *
22 TELUGU	0C00-0C7F 200C, 200D	57, 58, 59 (These collection numbers shall not be used, see Note 2.)		
23 KANNADA	0C80-0CFF 200C, 200D	60	CJK UNIFIED IDEOGRAPHS	4E00-9FFF
24 MALAYALAM	0D00-0D7F 200C, 200D	61	PRIVATE USE AREA	E000-F8FF
25 THAI	0E00-0E7F	62	CJK COMPATIBILITY IDEOGRAPHS	F900-FAFF
26 LAO	0E80-0EFF	63	(Collection specified as union of other collections)	
27 BASIC GEORGIAN	10D0-10FF	64	ARABIC PRESENTATION FORMS-A	FB50-FDCF FDF0-FDFF
28 GEORGIAN EXTENDED	10A0-10CF	65	COMBINING HALF MARKS	FE20-FE2F
29 HANGUL JAMO	1100-11FF *	66	CJK COMPATIBILITY FORMS	FE30-FE4F *
30 LATIN EXTENDED ADDITIONAL	1E00-1EFF *	67	SMALL FORM VARIANTS	FE50-FE6F
31 GREEK EXTENDED	1F00-1FFF	68	ARABIC PRESENTATION FORMS-B	FE70-FEFE
32 GENERAL PUNCTUATION	2000-206F	69	HALFWIDTH AND FULLWIDTH FORMS	FF00-FFEF
33 SUPERSCRIPTS AND SUBSCRIPTS	2070-209F	70	SPECIALS	FFFO-FFFD
34 CURRENCY SYMBOLS	20A0-20CF	71	HANGUL SYLLABLES	AC00-D7A3 *
		72	BASIC TIBETAN	0F00-0FBF
		73	ETHIOPIC	1200-137F

ISO/IEC 10646:2014 (E)

74	UNIFIED CANADIAN ABORIGINAL SYLLABICS	1400-167F *	116	PHONETIC EXTENSIONS SUPPLEMENT	1D80-1DBF *
75	CHEROKEE	13A0-13FF	117	COMBINING DIACRITICAL MARKS SUPPLEMENT	1DC0-1DFF
76	YI SYLLABLES	A000-A48F	118	GLAGOLITIC	2C00-2C5F
77	YI RADICALS	A490-A4CF	119	COPTIC	03E2-03EF 2C80-2CFF
78	KANGXI RADICALS	2F00-2FDF	120	GEORGIAN SUPPLEMENT	2D00-2D2F
79	CJK RADICALS SUPPLEMENT	2E80-2EFF	121	TIFINAGH	2D30-2D7F
80	BRAILLE PATTERNS	2800-28FF	122	ETHIOPIC EXTENDED	2D80-2DDF
81	CJK UNIFIED IDEOGRAPHS EXTENSION A	3400-4DBF FA1F, FA23	123	SUPPLEMENTAL PUNCTUATION	2E00-2E7F
82	OGHAM	1680-169F	124	CJK STROKES	31C0-31EF
83	RUNIC	16A0-16FF	125	MODIFIER TONE LETTERS	A700-A71F *
84	SINHALA	0D80-0DFF	126	SYLOTI NAGRI	A800-A82F
85	SYRIAC	0700-074F	127	VERTICAL FORMS	FE10-FE1F
86	THAANA	0780-07BF	128	NKO	07C0-07FF
87	BASIC MYANMAR	1000-104F * 200C, 200D	129	BALINESE	1B00-1B7F
88	KHMER	1780-17FF 200C, 200D	130	LATIN EXTENDED-C	2C60-2C7F *
89	MONGOLIAN	1800-18AF	131	LATIN EXTENDED-D	A720-A7FF
90	EXTENDED MYANMAR	1050-109F *	132	PHAGS-PA	A840-A87F
91	TIBETAN	0F00-0FFF	133	SUNDANESE	1B80-1BBF *
92	CYRILLIC SUPPLEMENT	0500-052F	134	LEPCHA	1C00-1C4F
93	TAGALOG	1700-171F	135	OL CHIKI	1C50-1C7F *
94	HANUNOO	1720-173F	136	VAI	A500-A63F
95	BUHID	1740-175F	137	SAURASHTRA	A880-A8DF
96	TAGBANWA	1760-177F	138	KAYAH LI	A900-A92F *
97	MISCELLANEOUS MATHEMATICAL SYMBOLS-A	27C0-27EF *	139	REJANG	A930-A95F
98	SUPPLEMENTAL ARROWS-A	27F0-27FF *	140	CYRILLIC EXTENDED-A	2DE0-2DFF *
99	SUPPLEMENTAL ARROWS-B	2900-297F *	141	CYRILLIC EXTENDED-B	A640-A69F
100	MISCELLANEOUS MATHEMATICAL SYMBOLS-B	2980-29FF *	142	CHAM	AA00-AA5F
101	SUPPLEMENTAL MATHEMATICAL OPERATORS	2A00-2AFF *	143	TAI THAM	1A20-1AAF
102	KATAKANA PHONETIC EXTENSIONS	31F0-31FF *	144	HANGUL JAMO EXTENDED-A	A960-A97F
103	VARIATION SELECTORS	FE00-FE0F *	145	TAI VIET	AA80-AADF
104	LTR ALPHABETIC PRESENTATION FORMS	FB00-FB1C	146	HANGUL JAMO EXTENDED-B	D7B0-D7FF
105	RTL ALPHABETIC PRESENTATION FORMS	FB1D-FB4F	147	SAMARITAN	0800-083F
106	LIMBU	1900-194F	148	UNIFIED CANADIAN ABORIGINAL SYLLABICS EXTENDED	18B0-18FF
107	TAI LE	1950-197F	149	VEDIC EXTENSIONS	1CD0-1CFF
108	KHMER SYMBOLS	19E0-19FF *	150	LISU	A4D0-A4FF *
109	PHONETIC EXTENSIONS	1D00-1D7F *	151	BAMUM	A6A0-A6FF
110	MISCELLANEOUS SYMBOLS AND ARROWS	2B00-2BFF	152	COMMON INDIC NUMBER FORMS	A830-A83F
111	YIJING HEXAGRAM SYMBOLS	4DC0-4DFF *	153	DEVANAGARI EXTENDED	A8E0-A8FF
112	ARABIC SUPPLEMENT	0750-077F *	154	JAVANESE	A980-A9DF
113	ETHIOPIC SUPPLEMENT	1380-139F	155	MYANMAR EXTENDED-A	AA60-AA7F *
114	NEW TAI LUE	1980-19DF	156	MEETEI MAYEK	ABC0-ABFF
115	BUGINESE	1A00-1A1F	157	MANDAIC	0840-085F
			158	BATAK	1BC0-1BFF
			159	ETHIOPIC EXTENDED-A	AB00-AB2F
			160	ARABIC EXTENDED-A	08A0-08FF
			161	SUNDANESE SUPPLEMENT	1CC0-1CCF
			162	MEETEI MAYEK EXTENSIONS	AAE0-AAFF

163	COMBINING DIACRITICAL MARKS EXTENDED	1AB0-1AFF	1044	PLAYING CARDS	1F0A0-1F0FF
164	MYANMAR EXTENDED-B	A9E0-A9FF	1045	MISCELLANEOUS SYMBOLS AND PICTOGRAPHS	1F300-1F5FF
165	LATIN EXTENDED-E	AB30-AB6F	1046	EMOTICONS	1F600-1F64F
1001	OLD ITALIC	10300-1032F	1047	TRANSPORT AND MAP SYMBOLS	1F680-1F6FF
1002	GOTHIC	10330-1034F	1048	ALCHEMICAL SYMBOLS	1F700-1F77F
1003	DESERET	10400-1044F *	1049	MEROITIC HIEROGLYPHS	10980-1099F *
1004	BYZANTINE MUSICAL SYMBOLS	1D000-1D0FF	1050	MEROITIC CURSIVE	109A0-109FF
1005	MUSICAL SYMBOLS	1D100-1D1FF	1051	SORA SOMPENG	110D0-110FF
1006	MATHEMATICAL ALPHANUMERIC SYMBOLS	1D400-1D7FF	1052	CHAKMA	11100-1114F
1007	LINEAR B SYLLABARY	10000-1007F	1053	SHARADA	11180-111DF
1008	LINEAR B IDEOGRAMS	10080-100FF	1054	TAKRI	11680-116CF
1009	AEGEAN NUMBERS	10100-1013F	1055	MIAO	16F00-16F9F
1010	UGARITIC	10380-1039F	1056	ARABIC MATHEMATICAL ALPHABETIC SYMBOLS	1EE00-1EEFF
1011	SHAVIAN	10450-1047F *	1057	COPTIC EPACT NUMBERS	102E0-102FF
1012	OSMANYA	10480-104AF	1058	ELBASAN	10500-1052F
1013	CYPRIOT SYLLABARY	10800-1083F	1059	LINEAR A	10600-1077F
1014	TAI XUAN JING SYMBOLS	1D300-1D35F	1060	PALMYRENE	10860-1087F *
1015	ANCIENT GREEK NUMBERS	10140-1018F	1061	NABATAEAN	10880-108AF
1016	OLD PERSIAN	103A0-103DF	1062	OLD NORTH ARABIAN	10A80-10A9F *
1017	KHAROSHTHI	10A00-10A5F	1063	MANICHAEAN	10AC0-10AFF
1018	ANCIENT GREEK MUSICAL NOTATION	1D200-1D24F	1064	SINHALA ARCHAIC NUMBERS	111E0-111FF
1019	PHOENICIAN	10900-1091F	1065	KHOJKI	11200-1124F
1020	CUNEIFORM	12000-123FF	1066	KHUDAWADI	112B0-112FF
1021	CUNEIFORM NUMBERS AND PUNCTUATION	12400-1247F	1067	TIRHUTA	11480-114DF
1022	COUNTING ROD NUMERALS	1D360-1D37F	1068	PAU CIN HAU	11AC0-11AFF
1023	PHAISTOS DISC	101D0-101FF	1069	MRO	16A40-16A 6F
1024	LYCIAN	10280-1029F	1070	BASSA VAH	16AD0-16AFF
1025	CARIAN	102A0-102DF	1071	DUPLOYAN	1BC00-1BC9F
1026	LYDIAN	10920-1093F	1072	SHORTHAND FORMAT CONTROLS	1BCA0-1BCAF
1027	ANCIENT SYMBOLS	10190-101CF	1073	ORNAMENTAL DINGBATS	1F650-1F67F *
1028	MAHJONG TILES	1F000-1F02F	1074	GEOMETRIC SHAPES EXTENDED	1F780-1F7FF
1029	DOMINO TILES	1F030-1F09F	1075	SUPPLEMENTAL ARROWS-C	1F800-1F8FF
1030	AVESTAN	10B00-10B3F	1076	OLD PERMIC	10350-1037F
1031	EGYPTIAN HIEROGLYPHS	13000-1342F	1077	CAUCASIAN ALBANIAN	10530-1056F
1032	IMPERIAL ARAMAIC	10840-1085F	1078	PSALTER PAHLAVI	10B80-10BAF
1033	OLD SOUTH ARABIAN	10A60-10A7F	1079	MAHAJANI	11150-1117F
1034	INSCRIPTIONAL PARTHIAN	10B40-10B5F	1080	GRANTHA	11300-1137F
1035	INSCRIPTIONAL PAHLAVI	10B60-10B7F	1081	SIDDHAM	11580-115FF
1036	OLD TURKIC	10C00-10C4F	1082	MODI	11600-1165F
1037	RUMI NUMERAL SYMBOLS	10E60-10E7F	1083	WARANG CITI	118A0-118FF
1038	KAITHI	11080-110CF	1084	PAHAUWH HMONG	16B00-16B8F
1039	ENCLOSED ALPHANUMERIC SUPPLEMENT	1F100-1F1FF	1085	MENDE KIKAKUI	1E800-1E8DF
1040	ENCLOSED IDEOGRAPHIC SUPPLEMENT	1F200-1F2FF	1086	HATRAN	108E0-108FF
1041	BRAHMI	11000-1107F	1087	OLD HUNGARIAN	10C80-10CFF
1042	KANA SUPPLEMENT	1B000-1B0FF	1088	MULTANI	11280-112AF
1043	BAMUM SUPPLEMENT	16800-16A3F	1089	AHOM	11700-1173F
			1090	EARLY DYNASTIC CUNEIFORM	12480-1254F
			1091	ANATOLIAN HIEROGLYPHS	14400-1467F
			1092	SUTTON SIGNWRITING	1D800-1DAAF

ISO/IEC 10646:2014 (E)

2001	CJK UNIFIED IDEOGRAPHS EXTENSION B	20000-2A6DF	2004	CJK UNIFIED IDEOGRAPHS EXTENSION D	2B740-2B81F
2002	CJK COMPATIBILITY IDEOGRAPHS SUPPLEMENT	2F800-2FA1F	2005	CJK UNIFIED IDEOGRAPHS EXTENSION E	2B820-2CEAF
2003	CJK UNIFIED IDEOGRAPHS EXTENSION C	2A700-2B73F	3001	TAGS	E0000-E007F
			3003	VARIATION SELECTORS SUPPLEMENT	E0100-E01EF *

The following collections specify characters used for alternate formats and script-specific formats. See Annex F for more information.

200	ZERO-WIDTH BOUNDARY INDICATORS	200B-200D FEFF
201	FORMAT SEPARATORS	2028-2029
202	BI-DIRECTIONAL FORMAT MARKS	200E-200F
203	BI-DIRECTIONAL FORMAT EMBEDDINGS	202A-202E
204	HANGUL FILL CHARACTERS	115F-1160 3164 FFA0
205	CHARACTER SHAPING SELECTORS	206A-206D
206	NUMERIC SHAPE SELECTORS	206E-206F
207	IDEOGRAPHIC DESCRIPTION CHARACTERS	2FF0-2FFF
208	CONTROL CHARACTERS	0000-001F 0007F-009F
3002	ALTERNATE FORMAT CHARACTERS	E0000-E0FFF

The following specify collections that represented the whole UCS when they were created

299	(This collection number shall not be used, see A.3.3) *	
301	BMP-AMD.7	see A.3.2 *
302	BMP SECOND EDITION	see A.3.4 *
303	UNICODE 3.1	see A.6.2 *
304	UNICODE 3.2	see A.6.3 *
305	UNICODE 4.0	see A.6.4 *
306	UNICODE 4.1	see A.6.5 *
307	UNICODE 5.0	see A.6.6 *
308	UNICODE 5.1	see A.6.7 *
309	UNICODE 5.2	see A.6.8 *
310	UNICODE 6.0	see A.6.9 *
311	UNICODE 6.1	see A.6.10 *
312	UNICODE 6.2	see A.6.11 *
313	UNICODE 6.3	see A.6.12 *
314	UNICODE 7.0	see A.6.13 *
340	COMBINED FIRST EDITION	see A.3.5 *
10646	UNICODE	0000-FDCF FDF0-FFFF 10000-1FFFFD 20000-2FFFFD 30000-3FFFFD 40000-4FFFFD 50000-5FFFFD 60000-6FFFFD 70000-7FFFFD 80000-8FFFFD 90000-9FFFFD A0000-AFFFFD B0000-BFFFFD C0000-CFFFFD D0000-DFFFFD E0000-EFFFFD F0000-FFFFD 100000-10FFFFD

NOTE 1 – The UNICODE collection incorporates all characters currently encoded in the standard

The following collections only contain CJK ideographs.

370	IICORE	see A.4.1 *
371	JIS2004 IDEOGRAPHICS EXTENSION	see A.4.2 *
372	JAPANESE IDEOGRAPHICS SUPPLEMENT	see A.4.3 *
380	CJK UNIFIED IDEOGRAPHS-2001	3400-4DB5 4E00-9FA5 FA0E-FA0F FA11 FA13-FA14 FA1F * FA21 FA23-FA24 FA27-FA29 20000-2A6D6

381	CJK COMPATIBILITY IDEOGRAPHS-2001	F900-FA0D FA10 FA12 FA15-FA1E FA20 FA22 FA25-FA26 * FA2A-FA6A 2F800-2FA1D
382	CJK UNIFIED IDEOGRAPHS-2005	Collection 380* 9FA6-9FBB
383	CJK COMPATIBILITY IDEOGRAPHS-2005	Collection 381 * FA70-FAD9
384	CJK UNIFIED IDEOGRAPHS-2007	Collection 382 * 9FBC-9FC3
385	CJK UNIFIED IDEOGRAPHS-2008	Collection 384 * 9FC4-9FC6 2A700-2B734
386	CJK COMPATIBILITY IDEOGRAPHS-2008	Collection 383 * FA6B-FA6D
387	CJK UNIFIED IDEOGRAPHS-2009	Collection 385 * 9FC7-9FCB 2B740-2B81D
388	CJK UNIFIED IDEOGRAPHS-2014	Collection 387 * 2B820-2CEA1

The following specify other collections, including extended collections.

270	COMBINING CHARACTERS	BMP characters specified in 4.14
271	(This collection number shall not be used, see Note 2) *	
281	MES-1	see A.5.2 *
282	MES-2	see A.5.3 *
283	MODERN EUROPEAN SCRIPTS	see A.5.4 *
284	CONTEMPORARY LITHUANIAN LETTERS	see A.5.5 *
285	BASIC JAPANESE	see A.5.6 *
286	JAPANESE NON IDEOGRAPHS EXTENSION	see A.5.7 *
287	COMMON JAPANESE	see A.5.8 *
288	MULTILINGUAL LATIN SUBSET	see A.5.9 *
300	BMP	0000-D7FF E000-FFFF
400	(This collection number shall not be used, see Note 3.)	
401	PRIVATE USE PLANES-0F-10	Planes 0F and 10
500	(This collection number shall not be used, see Note 3.)	
1000	SMP	10000-1FFFFD
1900	SMP COMBINING CHARACTERS	SMP characters specified in 4.14
2000	SIP	20000-2FFFFD
3000	SSP	E0000-EFFFFD

The following specify collections which are the union of particular collections defined above.

63	ALPHABETIC PRESENTATION FORMS	Collections 104-105
250	GENERAL FORMAT CHARACTERS	Collections 200-203
251	SCRIPT-SPECIFIC FORMAT CHARACTERS	Collections 204-206
4000	UCS PART-2	Collections 1000, 2000, 3000

NOTE 2 – Collections numbered 57, 58, and 59 were specified in the First Edition of ISO/IEC 10646-1 but have now been deleted. Collections numbered 400 and 500 were specified in the First and Second Editions of ISO/IEC 10646-1 but have now been deleted. The collection numbered 271 was specified in the first edition of this International Standard but has now been deleted.

NOTE 3 – The principal terms (keywords) used in the collection names shown above are listed below in alphabetical order. The entry for a term shows the collection number of every collection whose name includes the term. These terms do not provide a complete cross-reference to all the collections where characters sharing a particular attribute, such as script name, may be found. Although most of the terms identify an attribute of the characters within the collection, some characters that possess that attribute may be present in other collections whose numbers do not appear in the entry for that term.

Aegean numbers	1009	Anatolian Hieroglyphs	1091
Ahom	1089	Ancient Greek	1015 1018
Alphabetic	63	Arabic	14 15 64 68 112 160 1056
Alphanumeric	43	Armenian	11

ISO/IEC 10646:2014 (E)

Arrows	38 98 99 110 1075	Kharoshthi	1017
Avestan	1030	Khmer	88 108
Balinese	129	Khojki	1065
Bamum	151 1043	Khudawadi	1066
Bassa Vah	1070	Lao	26
Batak	158	Latin	1 2 3 4 30 130 131 165
Bengali	17	Lepcha	134
Bidirectional	202 203	Letter	36 55 1039 1040
Block elements	45	Limbu	106
BMP	300 301 302 (299)	Linear A	1059
Box drawing	44	Linear B ideograms	1008
Bopomofo	52	Linear B syllabary	1007
Brahmi	1041	Lisu	150
Braille patterns	80	Lycian	1024
Buginese	115	Lydian	1026
Buhid	95	Mahajani	1079
Byzantine musical symbols	1004	Malayalam	24
Canadian Aboriginal	74 148	Mandaic	157
Carian	1025	Manichaean	1063
Caucasian Albanian	1077	Mathematical alphanumeric symbols	1006 1056
Chakma	1052	Mathematical operators	39 101
Cham	142	Mathematical symbols	97 100
Cherokee	75	Meetei Mayek	156 162
CJK	49 54 55 56 60 62 66 78	Mende Kikakui	1085
	81 124 2001 2002	Meroitic	1049 1050
Combining	7 35 65 117 270 271	MES	281 282
Compatibility	53 56 62 66	Miao	1055
Control pictures	41	Modi	1082
Coptic	9 119 1057	Mongolian	89
Counting Rod numerals	1022	Months	55
Cuneiform	1020 1021 1090	Mro	1069
Currency	34	Multani	1088
Cypriot syllabary	1013	Musical notation	1018
Cyrillic	10 92 140 141	Musical symbols	1004 1005
Deseret	1003	Myanmar	87 90 155 164
Devanagari	16 153	Nabataean	1061
Diacritical marks	7 35 117 163	New Tai Lue	114
Dingbats	48 1073	NKo	128
Duployan	1071	Number	37 152 1009 1015
Elbasan	1058	Ogham	82
Enclosed	43 55	Ol Chiki	135
Egyptian hieroglyphs	1031	Old Hungarian	1087
Ethiopic	73 113 122 159	Old Italic	1001
Format	201 202 203 250 251	Old North Arabian	1062
	1072	Old Permic	1076
Fullwidth	69	Old Persian	1016
Game tiles	1028 1029	Old South Arabian	1033
Geometric shapes	46 1074	Old Turkic	1036
Georgian	27 28 120	Optical character recognition	42
Glagolitic	118	Oriya	20
Gothic	1002	Osmanya	1012
Grantha	1080	Pahawh Hmong	1084
Greek	8 9 31	Palmyrene	1060
Gujarati	19	Pau Cin Hau	1068
Gurmukhi	18	Phags-pa	132
Half (marks, width)	65 69	Phaistos Disc	1023
Hangul	29 53 71 144 146 204	Phoenician	1019
Hanunoo	94	Phonetic extensions	109 116
Hatran	1086	Presentation forms	63 64 68 104 105
Hebrew	12 13	Private use	61 401
Hiragana	50 1042	Psalter Pahlavi	1078
Ideographs	60 62 81 207 380-388	Punctuation	32 49 123
Imperial Aramaic	1032	Radicals	77 78 79
Inscriptional Pahlavi	1035	Rejang	139
Inscriptional Parthian	1034	Rumi numeral symbols	1037
IPA extensions	5	Runic	83
Jamo	29 53 144 146	Samaritan	147
Javanese	154	Saurashtra	137
Kaithi	1038	Shape, shaping	205 206
Kangxi	78	Sharada	1053
Kannada	23	Shavian	1011
Katakana	51 102 1042		
Kayah Li	138		

Siddham	1081	Tail Le	107
Sinhala	84 1064	Takri	1054
Small form	67	Tamil	21
Sora Sompeng	1051	Technical	40
Spacing modifier	6 125	Telugu	22
Specials	70	Thaana	86
Strokes	124	Thai	25
Subscripts, superscripts	33	Tibetan	72 91
Sundanese	133 161	Tifinagh	121
Sutton SignWriting	1092	Tirhuta	1067
Syllables, syllabics	71 74 76	Ugaritic	1010
Syloti Nagri	126	Unicode	303 304 305 306 307 308
Symbols	9 34 35 36 47 49 97 100		309 310 10646
	1027 1044 1045 1046	Vai	136
	1047 1048	Variation selectors	103 3003
Syriac	85	Vedic	149
Tagalog	93	Vertical form	127
Tagbanwa	96	Warang Citi	1083
Tags	3001	Yi	76 77
Tai Tham	143	Yijing hexagram symbols	111
Tai Viet	145	Zero-width	200
Tai Xuan Jing symbols	1014		

A.2 Blocks lists

A.2.1 Blocks in the BMP

The following blocks are specified in the Basic Multilingual Plane. They are ordered by code point.

Block name	from	to	
BASIC LATIN	0020-007E		ETHIOPIC SUPPLEMENT 1380-139F
LATIN-1 SUPPLEMENT	00A0-00FF		CHEROKEE 13A0-13FF
LATIN EXTENDED-A	0100-017F		UNIFIED CANADIAN ABORIGINAL SYLLABICS 1400-167F
LATIN EXTENDED-B	0180-024F		OGHAM 1680-169F
IPA EXTENSIONS	0250-02AF		RUNIC 16A0-16FF
SPACING MODIFIER LETTERS	02B0-02FF		TAGALOG 1700-171F
COMBINING DIACRITICAL MARKS	0300-036F		HANUNOO 1720-173F
GREEK AND COPTIC	0370-03FF		BUHID 1740-175F
CYRILLIC	0400-04FF		TAGBANWA 1760-177F
CYRILLIC SUPPLEMENT	0500-052F		KHMER 1780-17FF
ARMENIAN	0530-058F		MONGOLIAN 1800-18AF
HEBREW	0590-05FF		UNIFIED CANADIAN ABORIGINAL SYLLABICS EXTENDED 18B0-18FF
ARABIC	0600-06FF		LIMBU 1900-194F
SYRIAC	0700-074F		TAI LE 1950-197F
ARABIC SUPPLEMENT	0750-077F		NEW TAI LUE (Xishuang Banna Dai) 1980-19DF
THAANA	0780-07BF		KHMER SYMBOLS 19E0-19FF
NKO	07C0-07FF		BUGINESE 1A00-1A1F
SAMARITAN	0800-083F		TAI THAM 1A20-1AAF
MANDAIC	0840-085F		COMBINING DIACRITICAL MARKS EXTENDED 1AB0-1AFF
ARABIC EXTENDED-A	08A0-08FF		BALINESE 1B00-1B7F
DEVANAGARI	0900-097F		SUNDANESE 1B80-1BBF
BENGALI	0980-09FF		BATAK 1BC0-1BFF
GURMUKHI	0A00-0A7F		LEPCHA 1C00-1C4F
GUJARATI	0A80-0AFF		OL CHIKI 1C50-1C7F
ORIYA	0B00-0B7F		SUNDANESE SUPPLEMENT 1CC0-1CCF
TAMIL	0B80-0BFF		VEDIC EXTENSIONS 1CD0-1CFF
TELUGU	0C00-0C7F		PHONETIC EXTENSIONS 1D00-1D7F
KANNADA	0C80-0CFF		PHONETIC EXTENSIONS SUPPLEMENT 1D80-1DBF
MALAYALAM	0D00-0D7F		COMBINING DIACRITICAL MARKS SUPPLEMENT 1DC0-1DFF
SINHALA	0D80-0DFF		LATIN EXTENDED ADDITIONAL 1E00-1EFF
THAI	0E00-0E7F		GREEK EXTENDED 1F00-1FFF
LAO	0E80-0EFF		GENERAL PUNCTUATION 2000-206F
TIBETAN	0F00-0FFF		SUPERSCRIPTS AND SUBSCRIPTS 2070-209F
MYANMAR	1000-109F		CURRENCY SYMBOLS 20A0-20CF
GEORGIAN	10A0-10FF		
HANGUL JAMO	1100-11FF		
ETHIOPIC	1200-137F		

ISO/IEC 10646:2014 (E)

COMBINING DIACRITICAL MARKS FOR SYMBOLS	20D0-20FF	KATAKANA PHONETIC EXTENSIONS	31F0-31FF
LETTERLIKE SYMBOLS	2100-214F	ENCLOSED CJK LETTERS AND MONTHS	3200-32FF
NUMBER FORMS	2150-218F	CJK COMPATIBILITY	3300-33FF
ARROWS	2190-21FF	CJK UNIFIED IDEOGRAPHS EXTENSION A	3400-4DBF
MATHEMATICAL OPERATORS	2200-22FF	YIJING HEXAGRAM SYMBOLS	4DC0-4DFF
MISCELLANEOUS TECHNICAL	2300-23FF	CJK UNIFIED IDEOGRAPHS	4E00-9FFF
CONTROL PICTURES	2400-243F	YI SYLLABLES	A000-A48F
OPTICAL CHARACTER RECOGNITION	2440-245F	YI RADICALS	A490-A4CF
ENCLOSED ALPHANUMERICS	2460-24FF	LISU	A4D0-A4FF
BOX DRAWING	2500-257F	VAI	A500-A63F
BLOCK ELEMENTS	2580-259F	CYRILLIC EXTENDED-B	A640-A69F
GEOMETRIC SHAPES	25A0-25FF	BAMUM	A6A0-A6FF
MISCELLANEOUS SYMBOLS	2600-26FF	MODIFIER TONE LETTERS	A700-A71F
DINGBATS	2700-27BF	LATIN EXTENDED-D	A720-A7FF
MISCELLANEOUS MATHEMATICAL SYMBOLS-A	27C0-27EF	SYLOTI NAGRI	A800-A82F
SUPPLEMENTAL ARROWS-A	27F0-27FF	COMMON INDIC NUMBER FORMS	A830-A83F
BRAILLE PATTERNS	2800-28FF	PHAGS-PA	A840-A87F
SUPPLEMENTAL ARROWS-B	2900-297F	SAURASHTRA	A880-A8DF
MISCELLANEOUS MATHEMATICAL SYMBOLS-B	2980-29FF	DEVANAGARI EXTENDED	A8E0-A8FF
SUPPLEMENTAL MATHEMATICAL OPERATORS	2A00-2AFF	KAYAH LI	A900-A92F
MISCELLANEOUS SYMBOLS AND ARROWS	2B00-2BFF	REJANG	A930-A95F
GLAGOLITIC	2C00-2C5F	HANGUL JAMO EXTENDED-A	A960-A97F
LATIN EXTENDED-C	2C60-2C7F	JAVANESE	A980-A9DF
COPTIC	2C80-2CFF	MYANMAR EXTENDED-B	A9E0-A9FF
GEORGIAN SUPPLEMENT	2D00-2D2F	CHAM	AA00-AA5F
TIFINAGH	2D30-2D7F	MYANMAR EXTENDED-A	AA60-AA7F
ETHIOPIA EXTENDED	2D80-2DDF	TAI VIET	AA80-AADF
CYRILLIC EXTENDED-A	2DE0-2DFF	MEETEI MAYEK EXTENSIONS	AAE0-AAFF
SUPPLEMENTAL PUNCTUATION	2E00-2E7F	ETHIOPIA EXTENDED-A	AB00-AB2F
CJK RADICALS SUPPLEMENT	2E80-2EFF	LATIN EXTENDED-E	AB30-AB6F
KANGXI RADICALS	2F00-2FDF	MEETEI MAYEK	ABC0-ABFF
IDEOGRAPHIC DESCRIPTION CHARACTERS	2FF0-2FFF	HANGUL SYLLABLES	AC00-D7A3
CJK SYMBOLS AND PUNCTUATION	3000-303F	HANGUL JAMO EXTENDED-B	D7B0-D7FF
HIRAGANA	3040-309F	PRIVATE USE AREA	E000-F8FF
KATAKANA	30A0-30FF	CJK COMPATIBILITY IDEOGRAPHS	F900-FAFF
BOPOMOFO	3100-312F	ALPHABETIC PRESENTATION FORMS	FB00-FB4F
HANGUL COMPATIBILITY JAMO	3130-318F	ARABIC PRESENTATION FORMS-A	FB50-FDFF
KANBUN (CJK miscellaneous)	3190-319F	VARIATION SELECTORS	FE00-FE0F
BOPOMOFO EXTENDED	31A0-31BF	VERTICAL FORMS	FE10-FE1F
CJK STROKES	31C0-31EF	COMBINING HALF MARKS	FE20-FE2F
		CJK COMPATIBILITY FORMS	FE30-FE4F
		SMALL FORM VARIANTS	FE50-FE6F
		ARABIC PRESENTATION FORMS-B	FE70-FEFE
		HALFWIDTH AND FULLWIDTH FORMS	FF00-FFEF
		SPECIALS	FFFO-FFFD

NOTE – The parenthetical annotation located in some block names is not part of these names.

A.2.2 Blocks in the SMP

The following blocks are specified in the Supplementary Multilingual Plane for scripts and symbols. They are ordered by code point.

Block name	from	to	
LINEAR B SYLLABARY	10000-1007F	OLD PERMIC	10350-1037F
LINEAR B IDEOGRAMS	10080-100FF	UGARITIC	10380-1039F
AEGEAN NUMBERS	10100-1013F	OLD PERSIAN	103A0-103DF
ANCIENT GREEK NUMBERS	10140-1018F	DESERET	10400-1044F
ANCIENT SYMBOLS	10190-101CF	SHAVIAN	10450-1047F
PHAISTOS DISC	101D0-101FF	OSMANYA	10480-104AF
LYCIAN	10280-1029F	ELBASAN	10500-1052F
CARIAN	102A0-102DF	CAUCASIAN ALBANIAN	10530-1056F
COPTIC EPACT NUMBERS	102E0-102FF	LINEAR A	10600-1077F
OLD ITALIC	10300-1032F	CYPRIOT SYLLABARY	10800-1083F
GOTHIC	10330-1034F	IMPERIAL ARAMAIC	10840-1085F
		PALMYRENE	10860-1087F

NABATAEAN	10880-108AF	EARLY DYNASTIC CUNEIFORM	12480-1254F
HATRAN	108E0-108FF	EGYPTIAN HIEROGLYPHS	13000-1342F
PHOENICIAN	10900-1091F	ANATOLIAN HIEROGLYPHS	14400-1467F
LYDIAN	10920-1093F	BAMUM SUPPLEMENT	16800-16A3F
MEROITIC HIEROGLYPHS	10980-1099F	MRO	16A40-16A6F
MEROITIC CURSIVE	109A0-109FF	BASSA VAH	16AD0-16AFF
KHAROSHTHI	10A00-10A5F	PAHAWH HMONG	16B00-16B8F
OLD SOUTH ARABIAN	10A60-10A7F	MIAO	16F00-16F9F
OLD NORTH ARABIAN	10A80-10A9F	KANA SUPPLEMENT	1B000-1B0FF
MANICHAEAN	10AC0-10AFF	DUPLOYAN	1BC00-1BC9F
AVESTAN	10B00-10B3F	SHORTHAND FORMAT CONTROL	1BCA0-1BCAF
INSCRIPTIONAL PARTHIAN	10B40-10B5F	BYZANTINE MUSICAL SYMBOLS	1D000-1D0FF
INSCRIPTIONAL PAHLAVI	10B60-10B7F	MUSICAL SYMBOLS	1D100-1D1FF
PSALTER PAHLAVI	10B80-10BAF	ANCIENT GREEK MUSICAL NOTATION	1D200-1D24F
OLD TURKIC	10C00-10C4F	TAI XUAN JING SYMBOLS	1D300-1D35F
OLD HUNGARIAN	10C80-10CFF	COUNTING ROD NUMERALS	1D360-1D37F
RUMI NUMERAL SYMBOLS	10E60-10E7F	MATHEMATICAL ALPHANUMERIC	
BRAHMI	11000-1107F	SYMBOLS	1D400-1D7FF
KAITHI	11080-110CF	SUTTON SIGNWRITING	1D800-1DAAF
SORA SOMPENG	110D0-110FF	MENDE KIKAKUI	1E800-1E8DF
CHAKMA	11100-1114F	ARABIC MATHEMATICAL ALPHABETICAL	
MAHAJANI	11150-1117F	SYMBOLS	1EE00-1EEFF
SHARADA	11180-111DF	MAHJONG TILES	1F000-1F02F
SINHALA ARCHAIC NUMBERS	111E0-111FF	DOMINO TILES	1F030-1F09F
KHOJKI	11200-1124F	PLAYING CARDS	1F0A0-1F0FF
MULTANI	11280-112AF	ENCLOSED ALPHANUMERIC	
KHUDAWADI	112B0-112FF	SUPPLEMENT	1F100-1F1FF
GRANTHA	11300-1137F	ENCLOSED IDEOGRAPHIC	
TIRHUTA	11480-114DF	SUPPLEMENT	1F200-1F2FF
SIDDHAM	11580-115FF	MISCELLANEOUS SYMBOLS AND	
MODI	11600-1165F	PICTOGRAPHS	1F300-1F5FF
TAKRI	11680-116CF	EMOTICONS	1F600-1F64F
AHOM	11700-1173F	ORNAMENTAL DINGBATS	1F650-1F67F
WARANG CITI	118A0-118FF	TRANSPORT AND MAP SYMBOLS	1F680-1F6FF
PAU CIN HAU	11AC0-11AFF	ALCHEMICAL SYMBOLS	1F700-1F77F
CUNEIFORM	12000-123FF	GEOMETRIC SHAPES EXTENDED	1F780-1F7FF
CUNEIFORM NUMBERS AND		SUPPLEMENTAL ARROWS-C	1F800-1F8FF
PUNCTUATION	12400-1247F		

A.2.3 Blocks in the SIP

The following blocks are specified in the Supplementary Ideographic Plane. They are ordered by code point.

Block name	from	to
CJK UNIFIED IDEOGRAPHS EXTENSION B	20000	2A6DF
CJK UNIFIED IDEOGRAPHS EXTENSION C	2A700	2B73F
CJK UNIFIED IDEOGRAPHS EXTENSION D	2B740	2B81F
CJK UNIFIED IDEOGRAPHS EXTENSION E	2B820	2CEAF
CJK COMPATIBILITY IDEOGRAPHS SUPPLEMENT	2F800	2FA1F

A.2.4 Blocks in the SSP

The following blocks are specified in the Supplementary Special-purpose Plane. They are ordered by code point.

Block name	from	to
TAGS	E0000	E007F
VARIATION SELECTORS SUPPLEMENT	E0100	E01EF

A.3 Fixed collections of the whole UCS (except Unicode collections)

A.3.1 General

The following fixed collections (see 4.26) contain the whole UCS assigned character content as it was when they were created. The Unicode collections are described in A.6.

ISO/IEC 10646:2014 (E)

A.3.2 301 BMP-AMD.7

The fixed collection 301 BMP-AMD.7 is specified below. It comprises only those coded characters that were in the BMP after amendments up to , but not after, AMD.7 were applied to the First Edition of ISO/IEC 10646-1. Accordingly the repertoire of this collection is not subject to change if new characters are added to the BMP by any subsequent amendments.

301 BMP-AMD.7 is specified by the following ranges of code points as indicated for each row or contiguous series of rows.

Plane 00

Row	Values within row		
00	20-7E A0-FF	0F	00-47 49-69 71-8B 90-95 97 99-AD B1-B7 B9
01	00-F5 FA-FF	10	A0-C5 D0-F6 FB
02	00-17 50-A8 B0-DE E0-E9	11	00-59 5F-A2 A8-F9
03	00-45 60-61 74-75 7A 7E 84-8A 8C 8E-A1 A3-CE D0-D6 DA DC DE E0 E2-F3	1E	00-9B A0-F9
04	01-0C 0E-4F 51-5C 5E-86 90-C4 C7-C8 CB-CC D0-EB EE-F5 F8-F9	1F	00-15 18-1D 20-45 48-4D 50-57 59 5B 5D 5F-7D 80-B4 B6-C4 C6-D3 D6-DB DD-EF F2-F4 F6-FE
05	31-56 59-5F 61-87 89 91-A1 A3-B9 BB-C4 D0-EA F0-F4	20	00-2E 30-46 6A-70 74-8E A0-AB D0-E1
06	0C 1B 1F 21-3A 40-52 60-6D 70-B7 BA-BE C0-CE D0-ED F0-F9	21	00-38 53-82 90-EA
09	01-03 05-39 3C-4D 50-54 58-70 81-83 85-8C 8F-90 93-A8 AA-B0 B2 B6-B9 BC BE-C4 C7-C8 CB-CD D7 DC-DD DF-E3 E6-FA	22	00-F1
0A	02 05-0A 0F-10 13-28 2A-30 32-33 35-36 38-39 3C 3E-42 47-48 4B-4D 59-5C 5E 66-74 81-83 85-8B 8D 8F-91 93-A8 AA-B0 B2-B3 B5-B9 BC-C5 C7-C9 CB-CD D0 E0 E6-EF	23	00 02-7A
0B	01-03 05-0C 0F-10 13-28 2A-30 32-33 36-39 3C-43 47-48 4B-4D 56-57 5C-5D 5F-61 66-70 82-83 85-8A 8E-90 92-95 99-9A 9C 9E-9F A3-A4 A8-AA AE-B5 B7-B9 BE-C2 C6-C8 CA-CD D7 E7-F2	24	00-24 40-4A 60-EA
0C	01-03 05-0C 0E-10 12-28 2A-33 35-39 3E-44 46-48 4A-4D 55-56 60-61 66-6F 82-83 85-8C 8E-90 92-A8 AA-B3 B5-B9 BE-C4 C6-C8 CA-CD D5-D6 DE E0-E1 E6-EF	25	00-95 A0-EF
0D	02-03 05-0C 0E-10 12-28 2A-39 3E-43 46-48 4A-4D 57 60-61 66-6F	26	00-13 1A-6F
0E	01-3A 3F-5B 81-82 84 87-88 8A 8D 94-97 99-9F A1-A3 A5 A7 AA-AB AD-B9 BB-BD C0-C4 C6 C8-CD D0-D9 DC-DD	27	01-04 06-09 0C-27 29-4B 4D 4F-52 56 58-5E 61-67 76-94 98-AF B1-BE
		30	00-37 3F 41-94 99-9E A1-FE
		31	05-2C 31-8E 90-9F
		32	00-1C 20-43 60-7B 7F-B0 C0-CB D0-FE
		33	00-76 7B-DD E0-FE
		4E-9F	4E00-9FA5
		AC-D7	AC00-D7A3
		E0-F8	E000-F8FF
		F9-FA	F900-FA2D
		FB	00-06 13-17 1E-36 38-3C 3E 40-41 43-44 46-B1 D3-FF
		FC	00-FF
		FD	00-3F 50-8F 92-C7 F0-FB
		FE	20-23 30-44 49-52 54-66 68-6B 70-72 74 76-FC FF
		FF	01-5E 61-BE C2-C7 CA-CF D2-D7 DA-DC E0-E6 E8-EE FD

A.3.3 299 BMP FIRST EDITION

The fixed collection 299 BMP FIRST EDITION has been reserved to identify all of the coded characters that were in the BMP in the First Edition of ISO/IEC 10646-1. This collection is not now in conformity with this International Standard.

NOTE – The specification of collection 299 BMP FIRST EDITION consisted of the specification of collection 301 BMP-AMD.7 except for the replacement of the corresponding entries in the list above with the entries shown below:

Row	Values within row
05	31-56 59-5F 61-87 89 B0-B9 BB-C3 D0-EA F0-F4
0F	[no values]
1E	00-9A A0-F9
20	00-2E 30-46 6A-70 74-8E A0-AA D0-E1
AC-D7	[no values]

and by including an additional entry:

Row	Values within row
34-4D	3400-4DFF

for the code point values of three collections (57, 58, 59) of coded characters which have been deleted from this International Standard since the First Edition of ISO/IEC 10646-1.

A.3.4 302 BMP SECOND EDITION

The fixed collection 302 BMP SECOND EDITION comprises only those coded characters that are in the BMP in the Second Edition of ISO/IEC 10646-1. The repertoire of this collection is not subject to change if new characters are added to the BMP by any subsequent amendments.

302 BMP SECOND EDITION is specified by the following ranges of code points as indicated for each row or contiguous series of rows.

Plane 00

Row	Values within row		
00	20-7E A0-FF	13	00-0E 10 12-15 18-1E 20-46 48-5A 61-7C A0-F4
01	00-FF	14-15	1401-15FF
02	00-1F 22-33 50-AD B0-EE	16	00-76 80-9C A0-F0
03	00-4E 60-62 74-75 7A 7E 84-8A 8C 8E-A1 A3-CE D0-D7 DA-F3	17	80-DC E0-E9
04	00-86 88-89 8C-C4 C7-C8 CB-CC D0-F5 F8-F9	18	00-0E 10-19 20-77 80-A9
05	31-56 59-5F 61-87 89-8A 91-A1 A3-B9 BB-C4 D0-EA F0-F4	1E	00-9B A0-F9
06	0C 1B 1F 21-3A 40-55 60-6D 70-ED F0-FE	1F	00-15 18-1D 20-45 48-4D 50-57 59 5B 5D 5F-7D 80-B4 B6-C4 C6-D3 D6-DB DD-EF F2-F4 F6-FE
07	00-0D 0F-2C 30-4A 80-B0	20	00-46 48-4D 6A-70 74-8E A0-AF D0-E3
09	01-03 05-39 3C-4D 50-54 58-70 81-83 85-8C 8F-90 93-A8 AA-B0 B2 B6-B9 BC BE-C4 C7-C8 CB-CD D7 DC-DD DF-E3 E6-FA	21	00-3A 53-83 90-F3
0A	02 05-0A 0F-10 13-28 2A-30 32-33 35-36 38-39 3C 3E-42 47-48 4B-4D 59-5C 5E 66-74 81-83 85-8B 8D 8F-91 93-A8 AA-B0 B2-B3 B5-B9 BC-C5 C7-C9 CB-CD D0 E0 E6-EF	22	00-F1
0B	01-03 05-0C 0F-10 13-28 2A-30 32-33 36-39 3C-43 47-48 4B-4D 56-57 5C-5D 5F-61 66-70 82-83 85-8A 8E-90 92-95 99-9A 9C 9E-9F A3-A4 A8-AA AE-B5 B7-B9 BE-C2 C6-C8 CA-CD D7 E7-F2	23	00-7B 7D-9A
0C	01-03 05-0C 0E-10 12-28 2A-33 35-39 3E-44 46-48 4A-4D 55-56 60-61 66-6F 82-83 85-8C 8E-90 92-A8 AA-B3 B5-B9 BE-C4 C6-C8 CA-CD D5-D6 DE E0-E1 E6-EF	24	00-26 40-4A 60-EA
0D	02-03 05-0C 0E-10 12-28 2A-39 3E-43 46-48 4A-4D 57 60-61 66-6F 82-83 85-96 9A-B1 B3-BB BD C0-C6 CA CF-D4 D6 D8-DF F2-F4	25	00-95 A0-F7
0E	01-3A 3F-5B 81-82 84 87-88 8A 8D 94-97 99-9F A1-A3 A5 A7 AA-AB AD-B9 BB-BD C0-C4 C6 C8-CD D0-D9 DC-DD	26	00-13 19-71
0F	00-47 49-6A 71-8B 90-97 99-BC BE-CC CF	27	01-04 06-09 0C-27 29-4B 4D 4F-52 56 58-5E 61-67 76-94 98-AF B1-BE
10	00-21 23-27 29-2A 2C-32 36-39 40-59 A0-C5 D0-F6 FB	28	00-FF
11	00-59 5F-A2 A8-F9	2E	80-99 9B-F3
12	00-06 08-46 48 4A-4D 50-56 58 5A-5D 60-86 88 8A-8D 90-AE B0 B2-B5 B8-BE C0 C2-C5 C8-CE D0-D6 D8-EE F0-FF	2F	00-D5 F0-FB
		30	00-3A 3E-3F 41-94 99-9E A1-FE
		31	05-2C 31-8E 90-B7
		32	00-1C 20-43 60-7B 7F-B0 C0-CB D0-FE
		33	00-76 7B-DD E0-FE
		34-4D	3400-4DB5
		4E-9F	4E00-9FA5
		A0-A3	A000-A3FF
		A4	00-8C 90-A1 A4-B3 B5-C0 C2-C4 C6
		AC-D7	AC00-D7A3
		E0-F8	E000-F8FF
		F9-FA	F900-FA2D
		FB	00-06 13-17 1D-36 38-3C 3E 40-41 43-44 46-B1 D3-FF
		FC	00-FF
		FD	00-3F 50-8F 92-C7 F0-FB
		FE	20-23 30-44 49-52 54-66 68-6B 70-72 74 76-FC FF
		FF	01-5E 61-BE C2-C7 CA-CF D2-D7 DA-DC E0-E6 E8-EE F9-FD

A.3.5 340 COMBINED FIRST EDITION

The fixed collection 340 COMBINED FIRST EDITION is specified below. It comprises only those coded characters that were in the First Edition of this International Standard and consists of collections from A.1 and A.3 and several ranges of code points. The collection list is arranged by planes as follows.

Plane 00

Collection number and name

302	BMP SECOND EDITION
98	SUPPLEMENTAL ARROWS-A
99	SUPPLEMENTAL ARROWS-B
100	MISCELLANEOUS MATHEMATICAL SYMBOLS-B
101	SUPPLEMENTAL MATHEMATICAL OPERATORS
102	KATAKANA PHONETIC EXTENSIONS

ISO/IEC 10646:2014 (E)

103 VARIATION SELECTORS
108 KHMER SYMBOLS
111 YIJING HEXAGRAM SYMBOLS

Row	Values within row		
02	20-21 34-36 AE-AF EF-FF	21	3B 3D-4B F4-FF
03	4F-57 5D-5F 63-6F D8-D9 F4-FB	22	F2-FF
04	8A-8B C5-C6 C9-CA CD-CE	23	7C 9B-D0
05	00-0F	24	EB-FF
06	00-03 0D-15 56-58 6E-6F EE-EF FF	25	96-9F F8-FF
07	2D-2F 4D-4F B1	26	14-17 72-7D 80-91 A0-A1
09	04 BD	27	68-75 D0-EB
0A	01 03 8C E1-E3 F1	2B	00-0D
0B	35 71 F3-FA	30	3B-3D 95-96 9F-A0 FF
0C	BC-BD	32	1D-1E 50-5F 7C-7D B1-BF CC-CF
10	F7-F8	33	77-7A DE-DF FF
17	00-0C 0E-14 20-36 40-53 60-6C 6E-70 72-73 DD F0-F9	A4	A2-A3 B4 C1 C5
19	00-1C 20-2B 30-3B 40 44-4F 50-6D 70-74	FA	30-6A
1D	00-6B	FD	FC-FD
20	47 4E-54 57 5F-63 71 B0-B1 E4-EA	FE	45-48 73
		FF	5F-60

Plane 01

Collection number and name

1003 DESERET
1011 SHAVIAN

Row	Values within row
00	00-0B 0D-26 28-3A 3C-3D 3F-4D 50-5D 80-FA
01	00-02 07-33 37-3F
03	80-9D 9F
04	80-9D A0-A9
08	00-05 08 0A-35 37-38 3C 3F
D0	00-F5
D1	00-26 2A-DD
D3	00-56
D4	00-54 56-9C 9E-9F A2 A5-A6 A9-AC AE-B9 BB BD-C3 C5-FF
D5	00-05 07-0A 0D-14 16-1C 1E-39 3B-3E 40-44 46 4A-50 52-FF
D6	00-A3 A8-FF
D7	00-C9 CE-FF

Plane 02

Row Values within row

00-A6 0000-A6D6
F8-FA F800-FA1D

Plane 0E

Collection number and name

3003 VARIATION SELECTORS SUPPLEMENT

Row Values within row

00 01 20-7F

Plane 0F

Row Values within row

00-FF 0000-FFFD

Plane 10

Row Values within row

00-FF 0000-FFFD

A.4 CJK collections

A.4.1 370 IICORE

The fixed collection 370 IICORE is the International Core subset of the CJK UNIFIED IDEOGRAPHS-2001 collection.

NOTE 1 – Given its large size (9 810 c haracters) and the large number of sparse ranges, the collection is not specified by code point ranges in this document but instead by a linked content.

The content linked is the same file as the one used for the source references for the CJK ideographs. The presence of an IICORE tag (IIICORE) for a given CJK ideograph indicates that the character is part of the IICORE collection. See 23.2 for further details and access to the linked content.

A.4.2 371 JIS2004 IDEOGRAPHICS EXTENSION

The fixed collection 371 JIS2004 IDEOGRAPHICS EXTENSION consists of all level 3 and level 4 CJK characters defined in JIS X 0213:2004.

NOTE 1 – Given its large size (3 695 c haracters) and the large number of sparse ranges, the collection is not specified by code point ranges in this document but instead by a linked content.

The content linked to is a plain text file, using ISO/IEC 646-IRV characters with CARRIAGE RETURN/LINE FEED as end of line mark, that specifies, after a 3-lines header, as many lines as characters in the collection; each containing the following information in fixed length field:

- BMP or SIP code point (0hhhh), (2hhhh), normative.

The format definition uses ‘h’ as a hexadecimal unit. Digits between parentheses appear as shown.

[Click on this highlighted text to access the reference file.](#)

NOTE 2 – The content is also available as a separate viewable file in the same file directory as this document. The file is named: “JIExt.txt”.

A.4.3 372 JAPANESE IDEOGRAPHICS SUPPLEMENT

The fixed collection 372 JAPANESE IDEOGRAPHICS SUPPLEMENT consists of all CJK characters defined in JIS X 0212:1990. It contains 5 801 characters.

NOTE – 2 742 characters are common between the collections 371 and 372.

The code points of this collection are identified by the J1 Kanji J sources in the Source Reference file for CJK Unified Ideographs (CJKU_SR.txt). See 23.1 for further details.

A.5 Other collections

A.5.1 General

The collections specified within A.5 address the referencing need of users community. Characters may be from different writing systems and may be coded in different planes. It includes collection for users community from Lithuania, Japan and Europe as a whole.

NOTE – The acronym MES used in collection names below indicates Multilingual European Subset.

A.5.2 281 MES-1

The fixed collection 281 MES-1 is specified by the following ranges of code points as indicated for each row.

Plane 00

Row	Values within row
00	20-7E A0-FF
01	00-13 16-2B 2E-4D 50-7E
02	C7 D8-DB DD
20	15 18-19 1C-1D AC
21	22 26 5B-5E 90-93

ISO/IEC 10646:2014 (E)

26 6A

A.5.3 282 MES-2

The fixed collection 282 MES-2 is specified by the following ranges of code points as indicated for each row.

Plane 00

Row	Values within row
00	20-7E A0-FF
01	00-7F 8F 92 B7 DE-EF FA-FF
02	18-1B 1E-1F 59 7C 92 BB-BD C6-C7 C9 D8-DD EE
03	74-75 7A 7E 84-8A 8C 8E-A1 A3-CE D7 DA-E1
04	00-5F 90-C4 C7-C8 CB-CC D0-EB EE-F5 F8-F9
1E	02-03 0A-0B 1E-1F 40-41 56-57 60-61 6A-6B 80-85 9B F2-F3
1F	00-15 18-1D 20-45 48-4D 50-57 59 5B 5D 5F-7D 80-B4 B6-C4 C6-D3 D6-DB DD-EF F2-F4 F6-FE
20	13-15 17-1E 20-22 26 30 32-33 39-3A 3C 3E 44 4A 7F 82 A3-A4 A7 AC AF
21	05 16 22 26 5B-5E 90-95 A8
22	00 02-03 06 08-09 0F 11-12 19-1A 1E-1F 27-2B 48 59 60-61 64-65 82-83 95 97
23	02 10 20-21 29-2A
25	00 02 0C 10 14 18 1C 24 2C 34 3C 50-6C 80 84 88 8C 90-93 A0 AC B2 BA BC C4 CA-CB D8-D9
26	3A-3C 40 42 60 63 65-66 6A-6B
FB	01-02
FF	FD

A.5.4 283 MODERN EUROPEAN SCRIPTS

The collection 283 MODERN EUROPEAN SCRIPTS is specified by the following collections:

Collection number and name		
1	BASIC LATIN	34 CURRENCY SYMBOLS
2	LATIN-1 SUPPLEMENT	35 COMBINING DIACRITICAL MARKS FOR SYMBOLS
3	LATIN EXTENDED-A	36 LETTERLIKE SYMBOLS
4	LATIN EXTENDED-B	37 NUMBER FORMS
5	IPA EXTENSIONS	38 ARROWS
6	SPACING MODIFIER LETTERS	39 MATHEMATICAL OPERATORS
7	COMBINING DIACRITICAL MARKS	40 MISCELLANEOUS TECHNICAL
8	BASIC GREEK	42 OPTICAL CHARACTER RECOGNITION
9	GREEK SYMBOLS AND COPTIC	44 BOX DRAWING
10	CYRILLIC	45 BLOCK ELEMENTS
11	ARMENIAN	46 GEOMETRIC SHAPES
27	BASIC GEORGIAN	47 MISCELLANEOUS SYMBOLS
30	LATIN EXTENDED ADDITIONAL	65 COMBINING HALF MARKS
31	GREEK EXTENDED	70 SPECIALS
32	GENERAL PUNCTUATION	92 CYRILLIC SUPPLEMENT
33	SUPERSCRIPTS AND SUBSCRIPTS	104 LTR ALPHABETIC PRESENTATION FORMS

A.5.5 284 CONTEMPORARY LITHUANIAN LETTERS

The fixed extended collection 284 CONTEMPORARY LITHUANIAN LETTERS is defined as follows.

Plane 00

Row	Values within row
00	41-50 52-56 59-5A 61-70 72-76 79-7A C0-C1 C3 C8-C9 CC-CD D1-D3 D5 D9-DA DD E0-E1 E3 E8-E9 F1-F3 F5 F9-FA FD
01	04-05 0C-0D 16-19 28 2E-2F 60-61 68-6B 72-73 7D-7E
1E	BC-BD F8-F9

UCS Sequence Identifiers

<0104, 0301> <0105, 0301> <0104, 0303> <0105, 0303> <0118, 0301> <0119, 0301> <0118, 0303> <0119, 0303> <0116, 0301> <0117, 0301> <0116, 0303> <0117, 0303> <0069, 0307, 0300> <0069, 0307, 0301> <0069, 0307, 0303> <012E, 0301> <012F, 0307, 0301> <012E, 0303> <012F, 0307, 0303> <004A, 0303> <006A, 0307, 0303> <004C, 0303> <006C, 0303> <004D, 0303> <006D, 0303> <0052, 0303> <0072, 0303> <0172, 0301> <0173, 0301> <0172, 0303> <0173, 0303> <016A, 0301> <016B, 0301> <016A, 0303> <016B, 0303>

A.5.6 285 BASIC JAPANESE

The fixed collection 285 BASIC JAPANESE is a core Japanese subset. Its 6 884 characters are identified by:

- All J0 Kanji J sources in the Source Reference file for CJK Unified Ideographs (CJKU_SR.txt). See 23.1 for further details.
- Ranges of code points arranged by planes:

Plane 00

Row	Values within row		
00	20-7E A2 A3 A5 A7-A8 AC B0-B1 B4 B6 D7 F7	22	00 02-03 07-08 0B 12 1A 1D-1E 20 27-2C 34-35 3D 52 60-61 66-67 6A-6B 82-83 86-87 A5
03	91-A1 A3-A9 B1-C1 C3-C9	23	12
04	01 10-4F 51	25	00-03 0C 0F-10 13-14 17-18 1B-1D 20 23-25
20	10 14 16 18-19 1C-1D 20-21 25-26 30 32-33 3B 3E		28 2B-2C 2F-30 33-34 37-38 3B-3C 3F 42 4B
21	03 2B 90-93 D2 D4	26	A0-A1 B2-B3 BC-BD C6-C7 CB CE-CF EF
		30	05-06 40 42 6A 6D 6F
			00-03 05-15 1C 41-93 9B-9E A1-F6 FB-FE

A.5.7 286 JAPANESE NON IDEOGRAPHICS EXTENSION

The fixed collection 286 JAPANESE NON IDEOGRAPHICS EXTENSION is a Japanese subset which completes JIS X 0213 non-ideographic repertoire in combination with either 285 BASIC JAPANESE or 287 COMMON JAPANESE. Its 631 characters are identified by the following ranges of code points arranged by planes:

Plane 00

Row	Values within row		
00	A0-A1 A4 A6 A9-AB AD-AF B2-B3 B7-D6 D8-F6 F8-FF	22	05 09 13 1F 25-26 2E 43 45 48 62 76-77 84-85 8A-8B 95-97 BF DA-DB
01	00-09 0C-0F 11-13 18-1D 24-25 27 2A-2B 34-35 39-3A 3D-3E 41-44 47-48 4B-4D 50-55 58-65 6A-71 79-7E 93 C2 CD-CE D0-D2 D4 D6 D8 DA DC F8-F9 FD	23	05-06 18 BE-CC CE
02	50-5A 5C 5E-61 64-68 6C-73 75 79-7B 7D-7E 81-84 88-8E 90-92 94-95 98 9D A1-A2 C7-C8 CC D0-D1 D8-D9 DB DD-DE E5-E9	24	23 60-73 D0-E9 EB-FE
03	00-04 06 08 0B-0C 0F 18-1A 1C-20 24-25 29-2A 2C 2F-30 34 39-3D 61 C2	25	B1 B6-B7 C0-C1 C9 D0-D3 E6
1E	3E-3F	26	00-03 0E 16-17 1E 60-69 6B-6C 6E
1F	70-73	27	13 56 76-7F
20	13 22 3C 3F 42 47-49 51 AC	29	34-35 BF FA-FB
21	0F 13 16 21 27 35 53-55 60-6B 70-7B 94 96-99 C4 E6-E9	30	16-19 1D 1F-20 33-35 3B-3D 94-96 9A 9F-A0 F7-FA FF
		31	F0-FF
		32	31-32 39 51-5F A4-A8 B1-BF D0-E3 E5 E9 EC-ED FA
		33	03 0D 14 18 22-23 26-27 2B 36 3B 49-4A 4D
		FE	51 57 7B-7E 8E-8F 9C-9E A1 C4 CB CD
		FF	45-46
			5F-60

A.5.8 287 COMMON JAPANESE

The fixed collection 287 COMMON JAPANESE is a core Japanese subset containing 7 493 characters. It includes a fixed collection from A.5 and several ranges of code points.

Planes 00

Collection number and name

285 BASIC JAPANESE

Plane 00

Row	Values within row		
20	15	4E	28 E1 FC
21	16 21 60-69 70-79	4F	00 03 39 56 8A 92 94 9A C9 CD FF
22	11 1F 25 2E BF	50	1E 22 40 42 46 70 94 D8 F4
24	60-73	51	4A 64 9D BE EC
30	1D 1F	52	15 9C A6 AF C0 DB
32	31-32 39 A4-A8	53	00 07 24 72 93 B2 DD
33	03 0D 14 18 22-23 26-27 2B 36 3B 49-4A 4D	54	8A 9C A9 FF
	51 57 7B-7E 8E-8F 9C-9E A1 C4 CD	55	86
		57	59 65 AC C7-C8

ISO/IEC 10646:2014 (E)

58	9E B2	7B	9E
59	0B 53 5B 5D 63 A4 BA	7D	48 5C A0 B7 D6
5B	56 C0 D8 EC	7E	52 8A
5C	1E A6 BA F5	7F	47 A1
5D	27 42 53 6D B8-B9 D0	83	01 62 7F C7 F6
5F	21 34 45 67 B7 DE	84	48 B4 DC
60	5D 85 8A D5 DE F2	85	53 59 6B B0
61	11 20 30 37 98	88	07 F5
62	13 A6	89	1C
63	F5	8A	12 37 79 A7 BE DF F6
64	60 9D CE	8B	53 7F
65	4E	8C	F0 F4
66	00 09 15 1E 24 2E 31 3B 57 59 65 73 99 A0	8D	12 76
	B2 BF FA-FB	8E	CF
67	0E 66 BB C0	90	67 DE
68	01 44 52 C8 CF	91	15 27 D7 DA DE E4-E5 ED-EE
69	68 98 E2	92	06 0A 10 39-3A 3C 40 4E 51 59 67 77-78 88
6A	30 46 6B 73 7E E2 E4		A7 D0 D3 D5 D7 D9 E0 E7 F9 FB FF
6B	D6	93	02 1D-1E 21 25 48 57 70 A4 C6 DE F8
6C	3F 5C 6F 86 DA	94	31 45 48
6D	04 6F 87 96 AC CF F2 F8 FC	95	92
6E	27 39 3C 5C BF	96	9D AF
6F	88 B5 F5	97	33 3B 43 4D 4F 51 55
70	05 07 28 85 AB BB	98	57 65
71	04 0F 46-47 5C C1 FE	99	27 9E
72	B1 BE	9A	4E D9 DC
73	24 77 BD C9 D2 D6 E3 F5	9B	72 75 8F B1 BB
74	07 26 29-2A 2E 62 89 9F	9C	00
75	01 2F 6F	9D	6B 70
76	82 9B-9C 9E A6	9E	19 D1
77	46	F9	29 DC
78	21 4E 64 7A	FA	0E-2D
79	30 94 9B	FF	01-5E 61-9F E0-E5
7A	D1 E7 EB		

A.5.9 288 MULTILINGUAL LATIN SUBSET

The fixed collection 288 MULTILINGUAL LATIN SUBSET is an international Latin subset. It is specified by the following ranges of code points as indicated for each row.

NOTE – The collection 288 MULTILINGUAL LATIN SUBSET does not provide an exhaustive coverage of all languages using Latin-based orthographies. It is referenced by ISO/IEC 9995-3:2010 Keyboard layouts for text and office systems – Part 3: Complementary layouts of the alphanumeric zone of the alphanumeric section.

Plane 00

Row	Values within row
00	20-7E A0-FF
01	00-80 8F 97 9A-9B 9D-A1 AF-B0 B5-B7 CD-DC DE-F0 F4-F5 F8-FF
02	00-1B 1E-20 22-23 26-33 3A-3E 41-44 46-49 4C-4F 59 68 72 75 7C 89 92 94 B7 B9-BC BE-C1 C7-C8 CC-CD D8-DB DD
03	00-04 06-11 13 15 1B 23-29 2D-2E 31-32 35 38 44 47-48 5C-61
1D	7D CD
1E	00-19 1C-2B 2E-73 76-99 9B 9E A0-F9
20	0C 11 13-15 18-1A 1C-1E 26 2F 32-33 39-3A 4A A5 AC
21	22 26 4D 5B-5E 90-93 9A-9B
22	12 15 60 64-65 6E-71
23	00
26	6A
2C	63 65-66
A7	88 8B-8C

A.6 Unicode collections

A.6.1 General

These collections correspond to various versions of the Unicode Standard. They include characters from the BMP as well as Supplementary planes.

NOTE – Unicode 2.0 corresponds to collection 301. Unicode 2.1 adds the code points 20AC EURO SIGN and FFFC OBJECT REPLACEMENT CHARACTER to the collection 301. Unicode 3.0 corresponds to collection 302.

A.6.2 303 UNICODE 3.1

The fixed collection 303 UNICODE 3.1 consists of collections from A.3 and several ranges of code points. The collection list is arranged by planes as follows.

Plane 00

Collection number and name

302 BMP SECOND EDITION

Row Values within row

03 F4-F5

Plane 01

Row Values within row

03 00-1E 20-23 30-4A

04 00-25 28-4D

D0 00-F5

D1 00-26 2A-DD

D4 00-54 56-9C 9E-9F A2 A5-A6 A9-AC AE-B9 BB
BD-C0 C2-C3 C5-FF

D5 00-05 07-0A 0D-14 16-1C 1E-39 3B-3E 40-44
46 4A-50 52-FF

D6 00-A3 A8-FF

D7 00-C9 CE-FF

Plane 02

Row Values within row

00-A6 0000-A6D6

F8-FA F800-FA1D

Plane 0E

Row Values within row

00 01 20-7F

Plane 0F

Row Values within row

00-FF 0000-FFFD

Plane 10

Row Values within row

00-FF 0000-FFFD

A.6.3 304 UNICODE 3.2

The fixed collection 304 UNICODE 3.2 consists of fixed collections from A.1 and several ranges of code points arranged by planes as follows.

Planes 00-10

Collection number and name

303 UNICODE 3.1

Plane 00

Collection number and name

98 SUPPLEMENTAL ARROWS-A

99 SUPPLEMENTAL ARROWS-B

100 MISCELLANEOUS MATHEMATICAL SYMBOLS-B

101 SUPPLEMENTAL MATHEMATICAL OPERATORS

102 KATAKANA PHONETIC EXTENSIONS

103 VARIATION SELECTORS

ISO/IEC 10646:2014 (E)

Row	Values within row		
02	20	23	7C 9B-CE
03	4F 63-6F D8-D9 F6	24	EB-FE
04	8A-8B C5-C6 C9-CA CD-CE	25	96-9F F8-FF
05	00-0F	26	16-17 72-7D 80-89
06	6E-6F	27	68-75 D0-EB
07	B1	30	3B-3D 95-96 9F-A0 FF
10	F7-F8	32	51-5F B1-BF
17	00-0C 0E-14 20-36 40-53 60-6C 6E-70 72-73	A4	A2-A3 B4 C1 C5
20	47 4E-52 57 5F-63 71 B0-B1 E4-EA	FA	30-6A
21	3D-4B F4-FF	FE	45-46 73
22	F2-FF	FF	5F-60

A.6.4 305 UNICODE 4.0

The fixed collection 305 UNICODE 4.0 is identical to the fixed collection 340 COMBINED FIRST EDITION.

A.6.5 306 UNICODE 4.1

The fixed collection 306 UNICODE 4.1 consists of a fixed collection from A.1 and several ranges of code points. The collection list is arranged by planes as follows.

Plane 00-10

Collection number and name

305 UNICODE 4.0

Plane 00

Row	Values within row		
02	37-41	21	3C 4C
03	58-5C FC-FF	23	D1-DB
04	F6-F7	26	18 7E-7F 92-9C A2-B1
05	A2 C5-C7	27	C0-C6
06	0B 1E 59-5E	2B	0E-13
07	50-6D	2C	00-2E 30-5E 80-EA F9-FF
09	7D CE	2D	00-25 30-65 6F 80-96 A0-A6 A8-AE B0-B6 B8-BE C0-C6 C8-CE D0-D6 D8-DE
0B	B6 E6	2E	00-17 1C-1D
0F	D0-D1	31	C0-CF
10	F9-FA FC	32	7E
12	07 47 87 AF CF EF	9F	A6-BB
13	0F 1F 47 5F-60 80-99	A7	00-16
19	80-A9 B0-C9 D0-D9 DE-DF	A8	00-2B
1A	00-1B 1E-1F	FA	70-D9
1D	6C-C3	FE	10-19
20	55-56 58-5E 90-94 B2-B5 EB		

Plane 01

Row	Values within row		
01	40-8A	0A	00-03 05-06 0C-13 15-17 19-33 38-3A 3F-47
03	A0-C3 C8-D5	D2	50-58
		D6	00-45
			A4-A5

A.6.6 307 UNICODE 5.0

The fixed collection 307 UNICODE 5.0 consists of a fixed collection from A.1 and several ranges of code points. The collection list is arranged by planes as follows.

Plane 00-10

Collection number and name

306 UNICODE 4.1

Plane 00

Row	Values within row		
02	42-4F	09	7B-7C 7E-7F
03	7B-7D	0C	E2-E3 F1-F2
04	CF FA-FF	1B	00-4B 50-7C
05	10-13 BA	1D	C4-CA FE-FF
07	C0-FA	20	EC-EF
		21	4D-4E 84

23 DC-E7
 26 B2
 27 C7-CA
 2B 14-1A 20-23

2C 60-6C 74-77
 A7 17-1A 20-21
 A8 40-77

Plane 01

Row Values within row

09 00-19 1F
 20-22 2000-22FF
 23 00-6E

24 00-62 70-73
 D3 60-71
 D7 CA-CB

A.6.7 308 UNICODE 5.1

The fixed collection 308 UNICODE 5.1 is arranged by planes as follows.

Plane 00

Row Values within row

00 20-7E A0-FF
 01-02 0100-02FF
 03 00-77 7A-7E 84-8A 8C 8E-A1 A3-FF
 04 00-FF
 05 00-23 31-56 59-5F 61-87 89-8A 91-C7 D0-EA F0-F4
 06 00-03 06-1B 1E-1F 21-5E 60-FF
 07 00-0D 0F-4A 4D-B1 C0-FA
 09 01-39 3C-4D 50-54 58-72 7B-7F 81-83 85-8C 8F-90 93-A8 AA-B0 B2 B6-B9 BC-C4 C7-C8 CB-CE D7 DC-DD DF-E3 E6-FA
 0A 01-03 05-0A 0F-10 13-28 2A-30 32-33 35-36 38-39 3C 3E-42 47-48 4B-4D 51 59-5C 5E 66-75 81-83 85-8D 8F-91 93-A8 AA-B0 B2-B3 B5-B9 BC-C5 C7-C9 CB-CD D0 E0-E3 E6-EF F1
 0B 01-03 05-0C 0F-10 13-28 2A-30 32-33 35-39 3C-44 47-48 4B-4D 56-57 5C-5D 5F-63 66-71 82-83 85-8A 8E-90 92-95 99-9A 9C 9E-9F A3-A4 A8-AA AE-B9 BE-C2 C6-C8 CA-CD D0 D7 E6-FA
 0C 01-03 05-0C 0E-10 12-28 2A-33 35-39 3D-44 46-48 4A-4D 55-56 58-59 60-63 66-6F 78-7F 82-83 85-8C 8E-90 92-A8 AA-B3 B5-B9 BC-C4 C6-C8 CA-CD D5-D6 DE E0-E3 E6-EF F1-F2
 0D 02-03 05-0C 0E-10 12-28 2A-39 3D-44 46-48 4A-4D 57 60-63 66-75 79-7F 82-83 85-96 9A-B1 B3-BB BD C0-C6 CA CF-D4 D6 D8-DF F2-F4
 0E 01-3A 3F-5B 81-82 84 87-88 8A 8D 94-97 99-9F A1-A3 A5 A7 AA-A8 AD-B9 BB-BD C0-C4 C6 C8-CD D0-D9 DC-DD
 0F 00-47 49-6C 71-8B 90-97 99-BC BE-CC CE-D4
 10 00-99 9E-C5 D0-FC
 11 00-59 5F-A2 A8-F9
 12 00-48 4A-4D 50-56 58 5A-5D 60-88 8A-8D 90-B0 B2-B5 B8-BE C0 C2-C5 C8-D6 D8-FF
 13 00-10 12-15 18-5A 5F-7C 80-99 A0-F4
 14-15 1401-15FF
 16 00-76 80-9C A0-F0
 17 00-0C 0E-14 20-36 40-53 60-6C 6E-70 72-73 80-DD E0-E9 F0-F9
 18 00-0E 10-19 20-77 80-AA
 19 00-1C 20-2B 30-3B 40 44-6D 70-74 80-A9 B0-C9 D0-D9 DE-FF
 1A 00-1B 1E-1F
 1B 00-4B 50-7C 80-AA AE-B9
 1C 00-37 3B-49 4D-7F
 1D 00-E6 FE-FF

1E 00-FF
 1F 00-15 18-1D 20-45 48-4D 50-57 59 5B 5D 5F-7D 80-B4 B6-C4 C6-D3 D6-DB DD-EF F2-F4 F6-FE
 20 00-64 6A-71 74-8E 90-94 A0-B5 D0-F0
 21 00-4F 53-88 90-FF
 22 00-FF
 23 00-E7
 24 00-26 40-4A 60-FF
 25 00-FF
 26 00-9D A0-BC C0-C3
 27 01-04 06-09 0C-27 29-4B 4D 4F-52 56 58-5E 61-94 98-AF B1-BE C0-CA CC D0-FF
 28-2A 2800-2AFF
 2B 00-4C 50-54
 2C 00-2E 30-5E 60-6F 71-7D 80-EA F9-FF
 2D 00-25 30-65 6F 80-96 A0-A6 A8-AE B0-B6 B8-BE C0-C6 C8-CE D0-D6 D8-DE E0-FF
 2E 00-30 80-99 9B-F3
 2F 00-D5 F0-FB
 30 00-3F 41-96 99-FF
 31 05-2D 31-8E 90-B7 C0-E3 F0-FF
 32 00-1E 20-43 50-FE
 33 00-FF
 34-4C 3400-4CFF
 4D 00-B5 C0-FF
 4E-9F 4E00-9FC3
 A0-A3 A000-A3FF
 A4 00-8C 90-C6
 A5 00-FF
 A6 00-2B 40-5F 62-73 7C-97
 A7 00-8C FB-FF
 A8 00-2B 40-77 80-C4 CE-D9
 A9 00-53 5F
 AA 00-36 40-4D 50-59 5C-5F
 AC-D7 AC00-D7A3
 E0-F8 E000-F8FF
 F9 00-FF
 FA 00-2D 30-6A 70-D9
 FB 00-06 13-17 1D-36 38-3C 3E 40-41 43-44 46-B1 D3-FF
 FC 00-FF
 FD 00-3F 50-8F 92-C7 F0-FD
 FE 00-19 20-26 30-52 54-66 68-6B 70-74 76-FC FF
 FF 01-BE C2-C7 CA-CF D2-D7 DA-DC E0-E6 E8-EE F9-FD

Plane 01

ISO/IEC 10646:2014 (E)

Row	Values within row		
00	00-0B 0D-26 28-3A 3C-3D 3F-4D 50-5D 80-FA	24	00-62 70-73
01	00-02 07-33 37-8A 90-9B D0-FD	D0	00-F5
02	80-9C A0-D0	D1	00-26 29-DD
03	00-1E 20-23 30-4A 80-9D 9F-C3 C8-D5	D2	00-45
04	00-9D A0-A9	D3	00-56 60-71
08	00-05 08 0A-35 37-38 3C 3F	D4	00-54 56-9C 9E-9F A2 A5-A6 A9-AC AE-B9 BB
09	00-19 1F-39 3F		BD-C3 C5-FF
0A	00-03 05-06 0C-13 15-17 19-33 38-3A 3F-47	D5	00-05 07-0A 0D-14 16-1C 1E-39 3B-3E 40-44
	50-58	D6	00-A5 A8-FF
20-22	2000-22FF	D7	00-CB CE-FF
23	00-6E	F0	00-2B 30-93

Plane 02

Row	Values within row
00-A6	0000-A6D6
F8-FA	F800-FA1D

Plane 0E

Row	Values within row
00	01 20-7F
01	00-EF

Plane 0F

Row	Values within row
00-FF	0000-FFFF

Plane 10

Row	Values within row
00-FF	0000-FFFF

NOTE – The collection 308 UNICODE 5.1 can also be determined by using another fixed collection from A.1 and several ranges of code points.

Plane 00-10

Collection number and name
308 UNICODE 5.0

Plane 00

Row	Values within row		
03	70-73 76-77 CF	20	64 F0
04	87	21	4F 85-88
05	14-23	26	9D B3-BC C0-C3
06	06-0A 16-1A 3B-3F	27	CC EC-EF
07	6E-7F	2B	1B-1F 24-4C 50-54
09	71-72	2C	6D-6F 71-73 78-7D
0A	51 75	2D	E0-FF
0B	44 62-63 D0	2E	18-1B 1E-30
0C	3D 58-59 62-63 78-7F	31	2D D0-E3
0D	3D 44 62-63 70-75 79-7F	9F	BC-C3
0F	6B-6C CE D2-D4	A5	00-FF
10	22 28 2B 33-35 3A-3F 5A-8A	A6	00-2B 40-5F 62-73 7C-97
18	AA	A7	1B-1F 22-8C FB-FF
1B	80-AA AE-B9	A8	80-C4 CE-D9
1C	00-37 3B-49 4D-7F	A9	00-53 5F
1D	CB-E6	AA	00-36 40-4D 50-59 5C-5F
1E	9C-9F FA-FF	FE	24-26

Plane 01

Row	Values within row		
01	90-9B D0-FD	D1	29
02	80-9C A0-D0	F0	00-2B 30-93
09	20-39 3F		

A.6.8 309 UNICODE 5.2

The fixed collection 309 UNICODE 5.2 is arranged by planes as follows.

Plane 00

Row	Values within row		
00	20-7E A0-FF	1E	00-FF
01-02	0100-02FF	1F	00-15 18-1D 20-45 48-4D 50-57 59 5B 5D 5F-7D 80-B4 B6-C4 C6-D3 D6-DB DD-EF F2-F4 F6-FE
03	00-77 7A-7E 84-8A 8C 8E-A1 A3-FF	20	00-64 6A-71 74-8E 90-94 A0-B8 D0-F0
04	00-FF	21	00-89 90-FF
05	00-25 31-56 59-5F 61-87 89-8A 91-C7 D0-EA F0-F4	22	00-FF
06	00-03 06-1B 1E-1F 21-5E 60-FF	23	00-E8
07	00-0D 0F-4A 4D-B1 C0-FA	24	00-26 40-4A 60-FF
08	00-2D 30-3E	25	00-FF
09	00-39 3C-4E 50-55 58-72 79-7F 81-83 85-8C 8F-90 93-A8 AA-B0 B2 B6-B9 BC-C4 C7-C8 CB-CE D7 DC-DD DF-E3 E6-FB	26	00-CD CF-E1 E3 E8-FF
0A	01-03 05-0A 0F-10 13-28 2A-30 32-33 35-36 38-39 3C 3E-42 47-48 4B-4D 51 59-5C 5E 66-75 81-83 85-8D 8F-91 93-A8 AA-B0 B2-B3 B5-B9 BC-C5 C7-C9 CB-CD D0 E0-E3 E6-EF F1	27	01-04 06-09 0C-27 29-4B 4D 4F-52 56-5E 61-94 98-AF B1-BE C0-CA CC D0-FF
0B	01-03 05-0C 0F-10 13-28 2A-30 32-33 35-39 3C-44 47-48 4B-4D 56-57 5C-5D 5F-63 66-71 82-83 85-8A 8E-90 92-95 99-9A 9C 9E-9F A3-A4 A8-AA AE-B9 BE-C2 C6-C8 CA-CD D0 D7 E6-FA	28-2A	2800-2AFF
0C	01-03 05-0C 0E-10 12-28 2A-33 35-39 3D-44 46-48 4A-4D 55-56 58-59 60-63 66-6F 78-7F 82-83 85-8C 8E-90 92-A8 AA-B3 B5-B9 BC-C4 C6-C8 CA-CD D5-D6 DE E0-E3 E6-EF F1-F2	2B	00-4C 50-59
0D	02-03 05-0C 0E-10 12-28 2A-39 3D-44 46-48 4A-4D 57 60-63 66-75 79-7F 82-83 85-96 9A-B1 B3-BB BD C0-C6 CA CF-D4 D6 D8-DF F2-F4	2C	00-2E 30-5E 60-7F 80-F1 F9-FF
0E	01-3A 3F-5B 81-82 84 87-88 8A 8D 94-97 99-9F A1-A3 A5 A7 AA-A8 AD-B9 BB-BD C0-C4 C6 C8-CD D0-D9 DC-DD	2D	00-25 30-65 6F 80-96 A0-A6 A8-AE B0-B6 B8-BE C0-C6 C8-CE D0-D6 D8-DE E0-FF
0F	00-47 49-6C 71-8B 90-97 99-BC BE-CC CE-D8	2E	00-31 80-99 9B-F3
10	00-C5 D0-FC	2F	00-D5 F0-FB
11	00-FF	30	00-3F 41-96 99-FF
12	00-48 4A-4D 50-56 58 5A-5D 60-88 8A-8D 90-B0 B2-B5 B8-BE C0 C2-C5 C8-D6 D8-FF	31	05-2D 31-8E 90-B7 C0-E3 F0-FF
13	00-10 12-15 18-5A 5F-7C 80-99 A0-F4	32	00-1E 20-FE
14-15	1400-15FF	33	00-FF
16	00-9C A0-F0	34-4C	3400-4CFF
17	00-0C 0E-14 20-36 40-53 60-6C 6E-70 72-73 80-DD E0-E9 F0-F9	4D	00-B5 C0-FF
18	00-0E 10-19 20-77 80-AA B0-F5	4E-9F	4E00-9FC6
19	00-1C 20-2B 30-3B 40 44-6D 70-74 80-AB B0-C9 D0-DA DE-FF	A0-A3	A000-A3FF
1A	00-1B 1E-5E 60-7C 7F-89 90-99 A0-AD	A4	00-8C 90-C6 D0-FF
1B	00-4B 50-7C 80-AA AE-B9	A5	00-FF
1C	00-37 3B-49 4D-7F D0-F2	A6	00-2B 40-5F 62-73 7C-97 A0-F7
1D	00-E6 FD-FF	A7	00-8C FB-FF
		A8	00-2B 30-39 40-77 80-C4 CE-D9 E0-FB
		A9	00-53 5F-7C 80-CD CF-D9 DE-DF
		AA	00-36 40-4D 50-59 5C-7B 80-C2 DB-DF
		AB	C0-ED F0-F9
		AC-D6	AC00-D6FF
		D7	00-A3 B0-C6 CB-FB
		E0-F8	E000-F8FF
		F9	00-FF
		FA	00-2D 30-6D 70-D9
		FB	00-06 13-17 1D-36 38-3C 3E 40-41 43-44 46-B1 D3-FF
		FC	00-FF
		FD	00-3F 50-8F 92-C7 F0-FD
		FE	00-19 20-26 30-52 54-66 68-6B 70-74 76-FC FF
		FF	01-BE C2-C7 CA-CF D2-D7 DA-DC E0-E6 E8-EE F9-FD

Plane 01

Row	Values within row		
00	00-0B 0D-26 28-3A 3C-3D 3F-4D 50-5D 80-FA	0A	00-03 05-06 0C-13 15-17 19-33 38-3A 3F-47 50-58 60-7F
01	00-02 07-33 37-8A 90-9B D0-FD	0B	00-35 39-55 58-72 78-7F
02	80-9C A0-D0	0C	00-48
03	00-1E 20-23 30-4A 80-9D 9F-C3 C8-D5	0E	60-7E
04	00-9D A0-A9	10	80-C1
08	00-05 08 0A-35 37-38 3C 3F-55 57-5F	20-22	2000-22FF
09	00-1B 1F-39 3F	23	00-6E
		24	00-62 70-73

ISO/IEC 10646:2014 (E)

30-34	3000-342E	D5	00-05 07-0A 0D-14 16-1C 1E-39 3B-3E 40-44 46 4A-50 52-FF
D0	00-F5	D6	00-A5 A8-FF
D1	00-26 29-DD	D7	00-CB CE-FF
D2	00-45	F0	00-2B 30-93
D3	00-56 60-71	F1	00-0A 10-2E 31 3D 3F 42 46 4A-4E 57 5F 79 7B-7C 7F 8A-8D 90
D4	00-54 56-9C 9E-9F A2 A5-A6 A9-AC AE-B9 BB BD-C3 C5-FF	F2	00 10-31 40-48

Plane 02

Row	Values within row
00-A6	0000-A6D6
A7-B7	A700-B734
F8-FA	F800-FA1D

Plane 0E

Row	Values within row
00	01 20-7F
01	00-EF

Plane 0F

Row	Values within row
00-FF	0000-FFFF

Plane 10

Row	Values within row
00-FF	0000-FFFF

NOTE – The collection 309 UNICODE 5.2 can also be determined by using another fixed collection from A.1 and several ranges of code points.

Plane 00-10

Collection number and name
308 UNICODE 5.1

Plane 00

Row	Values within row		
05	24-25	26	9E-9F BD-BF C4-CD CF-E1 E3 E8-FF
08	00-2D 30-3E	27	57
09	00 4E 55 79-7A FB	2B	55-59
0F	D5-D8	2C	70 7E-7F EB-F1
10	9A-9D	2D	E0-FF
11	5A-5E A3-A7 FA-FF	2E	31
14	00	32	44-4F
16	77-7F	9F	C4-C6
18	B0-F5	A4	D0-FF
19	AA-AB DA	A6	A0-F7
1A	20-5E 60-7C 7F-89 90-99 A0-AD	A8	30-39 E0-FB
1C	D0-F2	A9	60-7C 80-CD CF-D9 DE-DF
1D	FD	AA	60-7B 80-C2 DB-DF
20	B6-B8	AB	C0-ED F0-F9
21	50-52 89	D7	B0-C6 CB-FB
23	E8	FA	6B-6D

Plane 01

Row	Values within row		
08	40-55 57-5F	10	80-C1
09	1A-1B	30-34	3000-342E
0A	60-7F	F1	00-0A 10-2E 31 3D 3F 42 46 4A-4E 57 5F 79 7B-7C 7F 8A-8D 90
0B	00-35 39-55 58-72 78-7F	F2	00 10-31 40-48
0C	00-48		
0E	60-7E		

Plane 02

Row Values within row
A7-B7 A700-B734

A.6.9 310 UNICODE 6.0

The fixed collection 310 UNICODE 6.0 is arranged by planes as follows.

Plane 00

Row	Values within row	1E	00-FF
00	20-7E A0-FF	1F	00-15 18-1D 20-45 48-4D 50-57 59 5B 5D 5F-7D 80-B4 B6-C4 C6-D3 D6-DB DD-EF F2-F4 F6-FE
01-02	0100-02FF		
03	00-77 7A-7E 84-8A 8C 8E-A1 A3-FF	20	00-64 6A-71 74-8E 90-9C A0-B9 D0-F0
04	00-FF	21	00-89 90-FF
05	00-27 31-56 59-5F 61-87 89-8A 91-C7 D0-EA F0-F4	22	00-FF
06	00-03 06-1B 1E-FF	23	00-F3
07	00-0D 0F-4A 4D-B1 C0-FA	24	00-26 40-4A 60-FF
08	00-2D 30-3E 40-5B 5E	25	00-FF
09	00-77 79-7F 81-83 85-8C 8F-90 93-A8 AA-B0 B2 B6-B9 BC-C4 C7-C8 CB-CE D7 DC-DD DF-E3 E6-FB	26	00-FF
0A	01-03 05-0A 0F-10 13-28 2A-30 32-33 35-36 38-39 3C 3E-42 47-48 4B-4D 51 59-5C 5E 66-75 81-83 85-8D 8F-91 93-A8 AA-B0 B2-B3 B5-B9 BC-C5 C7-C9 CB-CD D0 E0-E3 E6-EF F1	27	01-CA CC CE-FF
0B	01-03 05-0C 0F-10 13-28 2A-30 32-33 35-39 3C-44 47-48 4B-4D 56-57 5C-5D 5F-63 66-77 82-83 85-8A 8E-90 92-95 99-9A 9C 9E-9F A3-A4 A8-AA AE-B9 BE-C2 C6-C8 CA-CD D0 D7 E6-FA	28-2A	2800-2AFF
0C	01-03 05-0C 0E-10 12-28 2A-33 35-39 3D-44 46-48 4A-4D 55-56 58-59 60-63 66-6F 78-7F 82-83 85-8C 8E-90 92-A8 AA-B3 B5-B9 BC-C4 C6-C8 CA-CD D5-D6 DE E0-E3 E6-EF F1-F2	2B	00-4C 50-59
0D	02-03 05-0C 0E-10 12-3A 3D-44 46-48 4A-4E 57 60-63 66-75 79-7F 82-83 85-96 9A-B1 B3-BB BD C0-C6 CA CF-D4 D6 D8-DF F2-F4	2C	00-2E 30-5E 60-7F 80-F1 F9-FF
0E	01-3A 3F-5B 81-82 84 87-88 8A 8D 94-97 99-9F A1-A3 A5 A7 AA-A8 AD-B9 BB-BD C0-C4 C6 C8-CD D0-D9 DC-DD	2D	00-25 30-65 6F-70 7F-96 A0-A6 A8-AE B0-B6 B8-BE C0-C6 C8-CE D0-D6 D8-DE E0-FF
0F	00-47 49-6C 71-97 99-BC BE-CC CE-DA	2E	00-31 80-99 9B-F3
10	00-C5 D0-FC	2F	00-D5 F0-FB
11	00-FF	30	00-3F 41-96 99-FF
12	00-48 4A-4D 50-56 58 5A-5D 60-88 8A-8D 90-B0 B2-B5 B8-BE C0 C2-C5 C8-D6 D8-FF	31	05-2D 31-8E 90-BA C0-E3 F0-FF
13	00-10 12-15 18-5A 5D-7C 80-99 A0-F4	32	00-1E 20-FE
14-15	1400-15FF	33	00-FF
16	00-9C A0-F0	34-4C	3400-4CFF
17	00-0C 0E-14 20-36 40-53 60-6C 6E-70 72-73 80-DD E0-E9 F0-F9	4D	00-B5 C0-FF
18	00-0E 10-19 20-77 80-AA B0-F5	4E-9F	4E00-9FCB
19	00-1C 20-2B 30-3B 40 44-6D 70-74 80-AB B0-C9 D0-DA DE-FF	A0-A3	A000-A3FF
1A	00-1B 1E-5E 60-7C 7F-89 90-99 A0-AD	A4	00-8C 90-C6 D0-FF
1B	00-4B 50-7C 80-AA AE-B9 C0-F3 FC-FF	A5	00-FF
1C	00-37 3B-49 4D-7F D0-F2	A6	00-2B 40-73 7C-97 A0-F7
1D	00-E6 FC-FF	A7	00-8E 90-91A0-A9 FA-FF
		A8	00-2B 30-39 40-77 80-C4 CE-D9 E0-FB
		A9	00-53 5F-7C 80-CD CF-D9 DE-DF
		AA	00-36 40-4D 50-59 5C-7B 80-C2 DB-DF
		AB	01-06 09-0E 11-16 20-26 28-2E C0-ED F0-F9
		AC-D6	AC00-D6FF
		D7	00-A3 B0-C6 CB-FB
		E0-F8	E000-F8FF
		F9	00-FF
		FA	00-2D 30-6D 70-D9
		FB	00-06 13-17 1D-36 38-3C 3E 40-41 43-44 46-C1 D3-FF
		FC	00-FF
		FD	00-3F 50-8F 92-C7 F0-FD
		FE	00-19 20-26 30-52 54-66 68-6B 70-74 76-FC FF
		FF	01-BE C2-C7 CA-CF D2-D7 DA-DC E0-E6 E8-EE F9-FD

Plane 01

Row	Values within row	09	00-1B 1F-39 3F
00	00-0B 0D-26 28-3A 3C-3D 3F-4D 50-5D 80-FA	0A	00-03 05-06 0C-13 15-17 19-33 38-3A 3F-47 50-58 60-7F
01	00-02 07-33 37-8A 90-9B D0-FD	0B	00-35 39-55 58-72 78-7F
02	80-9C A0-D0	0C	00-48
03	00-1E 20-23 30-4A 80-9D 9F-C3 C8-D5	0E	60-7E
04	00-9D A0-A9	10	00-4D 52-6F 80-C1
08	00-05 08 0A-35 37-38 3C 3F-55 57-5F		

ISO/IEC 10646:2014 (E)

20-22	2000-22FF	D5	00-05 07-0A 0D-14 16-1C 1E-39 3B-3E 40-44
23	00-6E		46 4A-50 52-FF
24	00-62 70-73	D6	00-A5 A8-FF
30-34	3000-342E	D7	00-CB CE-FF
68-6A	6800-6A38	F0	00-2B 30-93 A0-AE B1-BE C1-CF D1-DF
B0	00-01	F1	00-0A 10-2E 30-69 70-8E 90-9A E6-FF
D0	00-F5	F2	00-02 10-3A 40-48 50-51
D1	00-26 29-DD	F3	00-20 30-35 37-7C 80-93 A0-C4 C6-CA E0-F0
D2	00-45	F4	00-3E 40 42-F7 F9-FC
D3	00-56 60-71	F5	00-3D 50-67 FB-FF
D4	00-54 56-9C 9E-9F A2 A5-A6 A9-AC AE-B9 BB	F6	01-10 12-14 16 18 1A 1C-1E 20-25 28-2B 2D
	BD-C3 C5-FF	F7	30-33 35-40 45-4F 80-C5
			00-73

Plane 02

Row	Values within row
00-A6	0000-A6D6
A7-B6	A700-B6FF
B7	00-34 40-FF
B8	00-1D
F8-FA	F800-FA1D

Plane 0E

Row	Values within row
00	01 20-7F
01	00-EF

Plane 0F

Row	Values within row
00-FF	0000-FFFF

Plane 10

Row	Values within row
00-FF	0000-FFFF

NOTE – The collection 310 UNICODE 6.0 can also be determined by using another fixed collection from A.1 and several ranges of code points. It contains the code point 20B9 INDIAN RUPEE SIGN which is added in this edition of this International Standard.

Plane 00-10

Collection number and name
309 UNICODE 5.2

Plane 00

Row	Values within row		
05	26-27	26	CE E2 E4-E7
08	40-5B 5E	27	05 0A-0B 28 4C 4E 53-55 5F-60 95-97 B0
09	3A-3B 4F 56-57 73-77		BF CE CF
0B	72-77	2D	70 7F
0D	29 3A 4E	30	97
0F	8C-8F D9-DA	31	B8-BA
13	5D-5E	9F	C7-CB
1B	C0-F3 FC-FF	A6	60-61
1D	FC	A7	8D-8E 90-91 A0-A9 FA
20	B9	AB	01-06 09-0E 11-16 20-26 28-2E
23	E9-F3	FB	B2-C1

Plane 01

Row	Values within row		
10	00-4D 52-6F	F1	30 32-3C 3E 40-41 43-45 47-49 4F-56 58-5E 60-69 70-78 7A 7D-7E 80-89 8E-8F
68-6A	6800-6A38		91-9A E6-FF
B0	00-01	F2	01-02 32-3A 50-51
F0	A0-AE B1-BE C1-CF D1-DF	F3	00-20 30-35 37-7C 80-93 A0-C4 C6-CA E0-F0
		F4	00-3E 40 42-F7 F9-FC

F5 00-3D 50-67 FB-FF
 F6 01-10 12-14 16 18 1A 1C-1E 20-25 28-2B
 2D 30-33 35-40 45-4F 80-C5

Plane 02

Row Values within row
 B7 40-FF

F7 00-73

B8 00-1D

A.6.10 311 UNICODE 6.1

The fixed collection 311 UNICODE 6.1 is arranged by planes as follows.

Plane 00

<u>Row</u>	<u>Values within row</u>		
00	20-7E A0-FF	1E	00-FF
01-02	0100-02FF	1F	00-15 18-1D 20-45 48-4D 50-57 59 5B 5D 5F-7D 80-B4 B6-C4 C6-D3 D6-DB DD-EF F2-F4 F6-FE
03	00-77 7A-7E 84-8A 8C 8E-A1 A3-FF	20	00-64 6A-71 74-8E 90-9C A0-B9 D0-F0
04	00-FF	21	00-89 90-FF
05	00-27 31-56 59-5F 61-87 89-8A 8F 91-C7 D0-EA F0-F4	22	00-FF
06	00-04 06-1B 1E-FF	23	00-F3
07	00-0D 0F-4A 4D-B1 C0-FA	24	00-26 40-4A 60-FF
08	00-2D 30-3E 40-5B 5E A0 A2-AC E4-FE	25	00-FF
09	00-77 79-7F 81-83 85-8C 8F-90 93-A8 AA-B0 B2 B6-B9 BC-C4 C7-C8 CB-CE D7 DC-DD DF-E3 E6-FB	26	00-FF
0A	01-03 05-0A 0F-10 13-28 2A-30 32-33 35-36 38-39 3C 3E-42 47-48 4B-4D 51 59-5C 5E 66-75 81-83 85-8D 8F-91 93-A8 AA-B0 B2-B3 B5-B9 BC-C5 C7-C9 CB-CD D0 E0-E3 E6-F1	27	01-FF
0B	01-03 05-0C 0F-10 13-28 2A-30 32-33 35-39 3C-44 47-48 4B-4D 56-57 5C-5D 5F-63 66-77 82-83 85-8A 8E-90 92-95 99-9A 9C 9E-9F A3-A4 A8-AA AE-B9 BE-C2 C6-C8 CA-CD D0 D7 E6-FA	28-2A	2800-2AFF
0C	01-03 05-0C 0E-10 12-28 2A-33 35-39 3D-44 46-48 4A-4D 55-56 58-59 60-63 66-6F 78-7F 82-83 85-8C 8E-90 92-A8 AA-B3 B5-B9 BC-C4 C6-C8 CA-CD D5-D6 DE E0-E3 E6-EF F1-F2	2B	00-4C 50-59
0D	02-03 05-0C 0E-10 12-3A 3D-44 46-48 4A-4E 57 60-63 66-75 79-7F 82-83 85-96 9A-B1 B3-BB BD C0-C6 CA CF-D4 D6 D8-DF F2-F4	2C	00-2E 30-5E 60-7F 80-F3 F9-FF
0E	01-3A 3F-5B 81-82 84 87-88 8A 8D 94-97 99-9F A1-A3 A5 A7 AA-A8 AD-B9 BB-BD C0-C4 C6 C8-CD D0-D9 DC-DF	2D	00-25 27 2D 30-67 6F-70 7F-96 A0-A6 A8-AE B0-B6 B8-BE C0-C6 C8-CE D0-D6 D8-DE E0-FF
0F	00-47 49-6C 71-97 99-BC BE-CC CE-DA	2E	00-3B 80-99 9B-F3
10	00-C5 C7 CD D0-FF	2F	00-D5 F0-FB
11	00-FF	30	00-3F 41-96 99-FF
12	00-48 4A-4D 50-56 58 5A-5D 60-88 8A-8D 90-B0 B2-B5 B8-BE C0 C2-C5 C8-D6 D8-FF	31	05-2D 31-8E 90-BA C0-E3 F0-FF
13	00-10 12-15 18-5A 5D-7C 80-99 A0-F4	32	00-1E 20-FE
14-15	1400-15FF	33	00-FF
16	00-9C A0-F0	34-4C	3400-4CFF
17	00-0C 0E-14 20-36 40-53 60-6C 6E-70 72-73 80-DD E0-E9 F0-F9	4D	00-B5 C0-FF
18	00-0E 10-19 20-77 80-AA B0-F5	4E-9F	4E00-9FCC
19	00-1C 20-2B 30-3B 40 44-6D 70-74 80-AB B0-C9 D0-DA DE-FF	A0-A3	A000-A3FF
1A	00-1B 1E-5E 60-7C 7F-89 90-99 A0-AD	A4	00-8C 90-C6 D0-FF
1B	00-4B 50-7C 80-F3 FC-FF	A5	00-FF
1C	00-37 3B-49 4D-7F C0-C7 D0-F6	A6	00-2B 40-97 9F-F7
1D	00-E6 FC-FF	A7	00-8E 90-93 A0-AA F8-FF
		A8	00-2B 30-39 40-77 80-C4 CE-D9 E0-FB
		A9	00-53 5F-7C 80-CD CF-D9 DE-DF
		AA	00-36 40-4D 50-59 5C-7B 80-C2 DB-F6
		AB	01-06 09-0E 11-16 20-26 28-2E C0-ED F0-F9
		AC-D6	AC00-D6FF
		D7	00-A3 B0-C6 CB-FB
		E0-F8	E000-F8FF
		F9	00-FF
		FA	00-6D 70-D9
		FB	00-06 13-17 1D-36 38-3C 3E 40-41 43-44 46-C1 D3-FF
		FC	00-FF
		FD	00-3F 50-8F 92-C7 F0-FD
		FE	00-19 20-26 30-52 54-66 68-6B 70-74 76-FC FF
		FF	01-BE C2-C7 CA-CF D2-D7 DA-DC E0-E6 E8-EE F9-FD

Plane 01

<u>Row</u>	<u>Values within row</u>		
00	00-0B 0D-26 28-3A 3C-3D 3F-4D 50-5D 80-FA	02	80-9C A0-D0
01	00-02 07-33 37-8A 90-9B D0-FD	03	00-1E 20-23 30-4A 80-9D 9F-C3 C8-D5
		04	00-9D A0-A9

ISO/IEC 10646:2014 (E)

08	00-05 08 0A-35 37-38 3C 3F-55 57-5F	D3	00-56 60-71
09	00-1B 1F-39 3F 80-B7 BE-BF	D4	00-54 56-9C 9E-9F A2 A5-A6 A9-AC AE-B9 BB
0A	00-03 05-06 0C-13 15-17 19-33 38-3A 3F-47		BD-C3 C5-FF
	50-58 60-7F	D5	00-05 07-0A 0D-14 16-1C 1E-39 3B-3E 40-44
0B	00-35 39-55 58-72 78-7F		46 4A-50 52-FF
0C	00-48	D6	00-A5 A8-FF
0E	60-7E	D7	00-CB CE-FF
10	00-4D 52-6F 80-C1 D0-E8 F0-F9	EE	00-03 05-1F 21-22 24 27 29-32 34-37 39 3B
11	00-34 36-43 80-C8 D0-D9		42 47 49 4B 4D-4F 51-52 54 57 59 5B 5D 5F
16	80-B7 C0-C9		61-62 64 67-6A 6C-72 74-77 79-7C 7E 80-89
20-22	2000-22FF		8B-9B A1-A3 A5-A9 AB-BB F0-F1
23	00-6E	F0	00-2B 30-93 A0-AE B1-BE C1-CF D1-DF
24	00-62 70-73	F1	00-0A 10-2E 30-6B 70-8E 90-9A E6-FF
30-34	3000-342E	F2	00-02 10-3A 40-48 50-51
68-6A	6800-6A38	F3	00-20 30-35 37-7C 80-93 A0-C4 C6-CA E0-F0
6F	00-44 50-7E 8F-9F	F4	00-3E 40 42-F7 F9-FC
B0	00-01	F5	00-3D 40-43 50-67 FB-FF
D0	00-F5	F6	00-40 45-4F 80-C5
D1	00-26 29-DD	F7	00-73
D2	00-45		

Plane 02

Row	Values within row
00-A6	0000-A6D6
A7-B6	A700-B6FF
B7	00-34 40-FF
B8	00-1D
F8-FA	F800-FA1D

Plane 0E

Row	Values within row
00	01 20-7F
01	00-EF

Plane 0F

Row	Values within row
00-FF	0000-FFFD

Plane 10

Row	Values within row
00-FF	0000-FFFD

NOTE – The collection 311 UNICODE 6.1 can also be determined by using another fixed collection from A.1 and several ranges of code points.

Plane 00-10

Collection number and name
310 UNICODE 6.0

Plane 00

Row	Values within row		
05	8F	2C	F2-F3
06	04	2D	27 2D 66-67
08	A0 A2-AC E4-FE	2E	32-3B
0A	F0	9F	CC
0E	DE-DF	A6	74-7B 9F
10	C7 CD FD-FF	A7	92-93 AA F8-F9
1B	AB-AD BA-BF	AA	E0-F6
1C	C0-C7 F3-F6	FA	2E-2F
27	CB CD		

Plane 01

Row	Values within row		
09	80-B7 BE-BF	11	00-34 36-43 80-C8 D0-D9
10	D0-E8 F0-F9	16	80-B7 C0-C9
		6F	00-44 50-7E 8F-9F

EE	00-03 05-1F 21-22 24 27 29-32 34-37 39 3B 42 47 49 4B 4D-4F 51-52 54 57 59 5B 5D 5F 61-62 64 67-6A 6C-72 74-77 79-7C 7E 80-89 8B-9B A1-A3 A5-A9 AB-BB FO-F1	F5	40-43
F1	6A-6B	F6	00 11 15 17 19 1B 1F 26-27 2C 2E-2F 34

A.6.11 312 UNICODE 6.2

The fixed collection 312 UNICODE 6.2 consists of a fixed collection from A.1 and a single code point. The collection list is arranged by planes as follows.

Plane 00-10Collection number and name

311 UNICODE 6.1

Plane 00

Row	Values within row
20	BA

A.6.12 313 UNICODE 6.3

The fixed collection 313 UNICODE 6.3 consists of a fixed collection from A.1 and six code points. The collection list is arranged by planes as follows.

Plane 00-10Collection number and name

311 UNICODE 6.1

Plane 00

Row	Values within row
06	1C
20	66-69 BA

A.6.13 314 UNICODE 7.0

The fixed collection 314 UNICODE 7.0 consists of a fixed collection from A.1 and six code points. The collection list is arranged by planes as follows.

Plane 00

Row	Values within row		
00	20-7E A0-FF	0D	01-03 05-0C 0E-10 12-3A 3D-44 46-48 4A-4E 57 60-63 66-75 79-7F 82-83 85-96 9A-B1 B3- BB BD C0-C6 CA CF-D4 D6 D8-DF E6-EF F2-F4
01-02	0100-02FF	0E	01-3A 3F-5B 81-82 84 87-88 8A 8D 94-97 99- 9F A1-A3 A5 A7 AA-A8 AD-B9 BB-BD C0-C4 C6 C8-CD D0-D9 DC-DF
03	00-77 7A-7F 84-8A 8C 8E-A1 A3-FF	0F	00-47 49-6C 71-97 99-BC BE-CC CE-DA
04	00-FF	10	00-C5 C7 CD D0-FF
05	00-2F 31-56 59-5F 61-87 89-8A 8D-8F 91-C7 D0-EA F0-F4	11	00-FF
06	00-1C 1E-FF	12	00-48 4A-4D 50-56 58 5A-5D 60-88 8A-8D 90-B0 B2-B5 B8-BE C0 C2-C5 C8-D6 D8-FF
07	00-0D 0F-4A 4D-B1 C0-FA	13	00-10 12-15 18-5A 5D-7C 80-99 A0-F4
08	00-2D 30-3E 40-5B 5E A0-B2 E4-FF	14-15	1400-15FF
09	00-83 85-8C 8F-90 93-A8 AA-B0 B2 B6-B9 BC- C4 C7-C8 CB-CE D7 DC-DD DF-E3 E6-FB	16	00-9C A0-F8
0A	01-03 05-0A 0F-10 13-28 2A-30 32-33 35-36 38-39 3C 3E-42 47-48 4B-4D 51 59-5C 5E 66- 75 81-83 85-8D 8F-91 93-A8 AA-B0 B2-B3 B5- B9 BC-C5 C7-C9 CB-CD D0 E0-E3 E6-F1	17	00-0C 0E-14 20-36 40-53 60-6C 6E-70 72-73 80-DD E0-E9 F0-F9
0B	01-03 05-0C 0F-10 13-28 2A-30 32-33 35-39 3C-44 47-48 4B-4D 56-57 5C-5D 5F-63 66-77 82-83 85-8A 8E-90 92-95 99-9A 9C 9E-9F A3- A4 A8-AA AE-B9 BE-C2 C6-C8 CA-CD D0 D7 E6-FA	18	00-0E 10-19 20-77 80-AA B0-F5
0C	00-03 05-0C 0E-10 12-28 2A-39 3D-44 46-48 4A-4D 55-56 58-59 60-63 66-6F 78-7F 81-83 85-8C 8E-90 92-A8 AA-B3 B5-B9 BC-C4 C6-C8 CA-CD D5-D6 DE E0-E3 E6-EF F1-F2	19	00-1E 20-2B 30-3B 40 44-6D 70-74 80-AB B0- C9 D0-DA DE-FF
		1A	00-1B 1E-5E 60-7C 7F-89 90-99 A0-AD B0-BE
		1B	00-4B 50-7C 80-F3 FC-FF
		1C	00-37 3B-49 4D-7F C0-C7 D0-F6 F8-F9
		1D	00-F5 FC-FF
		1E	00-FF

ISO/IEC 10646:2014 (E)

1F	00-15 18-1D 20-45 48-4D 50-57 59 5B 5D 5F-7D 80-B4 B6-C4 C6-D3 D6-DB DD-EF F2-F4 F6-FE	4E-9F	4E00-9FCC
20	00-64 66-71 74-8E 90-9C A0-BD D0-F0	A0-A3	A000-A3FF
21	00-89 90-FF	A4	00-8C 90-C6 D0-FF
22	00-FF	A5	00-FF
23	00-FA	A6	00-2B 40-9D 9F-F7
24	00-26 40-4A 60-FF	A7	00-8E 90-AD B0-B1 F7-FF
25	00-FF	A8	00-2B 30-39 40-77 80-C4 CE-D9 E0-FB
26	00-FF	A9	00-53 5F-7C 80-CD CF-D9 DE-FE
27	00-FF	AA	00-36 40-4D 50-59 5C-C2 DB-F6
28-2A	2800-2AFF	AB	01-06 09-0E 11-16 20-26 28-2E 30-5F C0-ED F0-F9
2B	00-73 76-95 98-B9 BD-C8 CA-D1	AC-D6	AC00-D6FF
2C	00-2E 30-5E 60-7F 80-F3 F9-FF	D7	00-A3 B0-C6 CB-FB
2D	00-25 27 2D 30-67 6F-70 7F-96 A0-A6 A8-AE B0-B6 B8-BE C0-C6 C8-CE D0-D6 D8-DE E0-FF	E0-F8	E000-F8FF
2E	00-42 80-99 9B-F3	F9	00-FF
2F	00-D5 F0-FB	FA	00-6D 70-D9
30	00-3F 41-96 99-FF	FB	00-06 13-17 1D-36 38-3C 3E 40-41 43-44 46-C1 D3-FF
31	05-2D 31-8E 90-BA C0-E3 F0-FF	FC	00-FF
32	00-1E 20-FE	FD	00-3F 50-8F 92-C7 F0-FD
33	00-FF	FE	00-19 20-2D 30-52 54-66 68-6B 70-74 76-FC FF
34-4C	3400-4CFF	FF	01-BE C2-C7 CA-CF D2-D7 DA-DC E0-E6 E8-EF F9-FD
4D	00-B5 C0-FF		

Plane 01

Row	Values within row		
00	00-0B 0D-26 28-3A 3C-3D 3F-4D 50-5D 80-FA	68-69	6800-69FF
01	00-02 07-33 37-8C 90-9B A0 D0-FD	6A	00-38 40-5E 60-69 6E-6F D0-ED F0-F5
02	80-9C A0-D0 E0-FB	6B	00-45 50-59 5B-61 63-77 7D-8F
03	00-1F 20-23 30-4A 50-7A 80-9D 9F-C3 C8-D5	6F	00-44 50-7E 8F-9F
04	00-9D A0-A9	B0	00-01
05	00-27 30-63 6F	BC	00-6A 70-7C 80-88 90-99 9C-A3
06	00-FF	D0	00-F5
07	00-36 40-55 60-67	D1	00-26 29-DD
08	00-05 08 0A-35 37-38 3C 3F-55 57-9E A7-AF	D2	00-45
09	00-1B 1F-39 3F 80-B7 BE-BF	D3	00-56 60-71
0A	00-03 05-06 0C-13 15-17 19-33 38-3A 3F-47 50-58 60-9F C0-E6 EB-F6	D4	00-54 56-9C 9E-9F A2 A5-A6 A9-AC AE-B9 BB BD-C3 C5-FF
0B	00-35 39-55 58-72 78-81 99-9C A9-AF	D5	00-05 07-0A 0D-14 16-1C 1E-39 3B-3E 40-44 46 4A-50 52-FF
0C	00-48	D6	00-A5 A8-FF
0E	60-7E	D7	00-CB CE-FF
10	00-4D 52-6F 7F-C1 D0-E8 F0-F9	E8	00-C4 C7-D6
11	00-34 36-43 50-76 80-C8 CD D0-DA E1-F4	EE	00-03 05-1F 21-22 24 27 29-32 34-37 39 3B 42 47 49 4B 4D-4F 51-52 54 57 59 5B 5D 5F 61-62 64 67-6A 6C-72 74-77 79-7C 7E 80-89 8B-9B A1-A3 A5-A9 AB-BB F0-F1
12	00-11 13-3D B0-EA F0-F9	F0	00-2B 30-93 A0-AE B1-BF C1-CF D1-F5
13	01-03 05-0C 0F-10 13-28 2A-30 32-33 35-39 3C-44 47-48 4B-4D 57 5D-63 66-6C 70-74	F1	00-0C 10-2E 30-6B 70-8E 90-9A E6-FF
14	80-C7 D0-D9	F2	00-02 10-3A 40-48 50-51
15	80-B5 B8-C9	F3	00-2C 30-35 37-7D 80-CE D4-F7
16	00-44 50-59 80-B7 C0-C9	F4	00-FE
18	A0-F2 FF	F5	00-4A 50-79 7B-A3 A5-FF
1A	C0-F8	F6	00-42 45-7F 80-CF E0-EC F0-F3
20-22	2000-22FF	F7	00-73 80-D4
23	00-98	F8	00-0B 10-47 50-59 60-87 90-AD
24	00-6E 70-74		
30-34	3000-342E		

Plane 02

Row	Values within row
00-A6	0000-A6D6
A7-B6	A700-B6FF
B7	00-34 40-FF
B8	00-1D
F8-FA	F800-FA1D

Plane 0E

<u>Row</u>	<u>Values within row</u>
00	01 20-7F
01	00-EF

Plane 0F

<u>Row</u>	<u>Values within row</u>
00-FF	0000-FFFD

Plane 10

<u>Row</u>	<u>Values within row</u>
00-FF	0000-FFFD

NOTE – The collection 314 UNICODE 7.0 can also be determined by using another fixed collection from A.1 and several ranges of code points.

Plane 00-10

<u>Collection number and name</u>
311 UNICODE 6.1

Plane 00

<u>Row</u>	<u>Values within row</u>		
03	7F	20	66-69 BB-BD
05	28-2F 8D-8E	23	F4-FA
06	05 1C	27	00
08	A1 AD-B2 FF	2B	4D-4F 5A-73 76-95 98-B9 BD-C8 CA-D1
09	78 80	2E	3C-42
0C	00 34 81	A6	98-9D
0D	01 E6-EF	A7	94-9F AB-AD B0-B1 F7
16	F1-F8	A9	E0-FE
19	1D-1E	AA	7C-7F
1A	B0-BE	AB	30-5F 64-65
1C	F8-F9	FE	27-2D
1D	E7-F5		

Plane 01

<u>Row</u>	<u>Values within row</u>		
01	8B-8C A0	16	00-44 50-59
02	E0-FB	18	A0-F2 FF
03	1F 50-7A	1A	C0-F8
05	00-27 30-63 6F	23	6F-98
06	00-FF	24	63-6E 74
07	00-36 40-55 60-67	6A	40-5E 60-69 6E-6F D0-ED F0-F5
08	60-9E A7-AF	6B	00-45 50-59 5B-61 63-77 7D-8F
0A	80-9F C0-E6 EB-F6	BC	00-6A 70-7C 80-88 90-99 9C-A3
0B	80-91 99-9C A9-AF	E8	00-C4 C7-D6
10	7F D0-E8 F0-F9	F0	BF E0-F5
11	50-76 CD DA E1-F4	F1	0B-0C
12	00-11 13-3D B0-EA F0-F9	F3	21-2C 36 7D 94-9F C5 CB-CE D4-DF F1-F7
13	01-03 05-0C 0F-10 13-28 2A-30 32-33 35-39 3C-44 47-48 4B-4D 57 5D-63 66- 6C 70-74	F4	3F 41 F8 FD-FE
14	80-C7 D0-D9	F5	3E-3F 44-4A 68-79 7B-A3 A5-FA
15	80-B5 B8-C9	F6	41-42 50-7F C6-CF E0-EC F0-F3
		F7	80-D4
		F8	00-0B 10-47 50-59 60-87 90-AD

Annex B
(normative)
List of combining characters

NOTE – Replaced by formal character class definition, see 4.14

Annex C
(normative)
Transformation format for planes 01 to 10 of the UCS (UTF-16)

NOTE – Incorporated in main body text, see UCS UTF-16 encoding form in 9 and UCS UTF-16 based encoding schemes in 10.

Annex D
(normative)
UCS Transformation Format 8 (UTF-8)

NOTE – Incorporated in main body text, see UCS UTF-8 encoding form in 9 and UCS UTF-8 encoding schemes in 10.

Annex E
(normative)
Mirrored characters in bidirectional context

NOTE – Replaced by formal character class definition for mirrored character, see 15.1.

Annex F (informative) Format characters

There is a special class of characters, called Format characters, the primary purpose of which is to affect the layout or processing of characters around them. With few exceptions, these characters do not have printable graphic symbols and, like the space characters, are represented in the character code charts by dotted boxes.

The function of most of these characters is to indicate the correct presentation of a code unit sequence. For any text processing other than presentation (such as sorting and searching), the format characters, except for ZWJ and ZWNJ described in F.1.1, can be ignored by filtering them out. The format characters are not intended to be used in conjunction with bidirectional control functions from ISO/IEC 6429.

F.1 General format characters

F.1.1 Hyphen boundary indicator

SOFT HYPHEN (00AD): SOFT HYPHEN (SHY) is a format character that indicates a preferred intra-word line-break opportunity. If the line is broken at that point, then whatever mechanism is appropriate for intra-word line-breaks should be invoked, just as if the line break had been triggered by another mechanism, such as a dictionary lookup. Depending on the language and the word, that may produce different visible results, such as:

- inserting a graphic symbol indicating the hyphenation and breaking the line after it,
- inserting a graphic symbol indicating the hyphenation, breaking the line after the symbol and changing spelling in the divided word parts,
- not showing any visible change and simply breaking the line at that point.

The inserted graphic symbol, if any, can take a wide variety of shapes, such as HYPHEN (2010), ARMENIAN HYPHEN (058A), MONGOLIAN TODO SOFT HYPHEN (1806), as appropriate for the situation.

When encoding text that includes explicit line breaking opportunities, including actual hyphenations, characters such as HYPHEN, ARMENIAN HYPHEN, and MONGOLIAN TODO SOFT HYPHEN may be used, depending on the language.

When a SOFT HYPHEN is inserted into a code unit sequence to encode a possible hyphenation point (for example: "tug{00AD}gumi"), the character representation remains otherwise unchanged. When encoding a code unit sequence that includes characters encoding hard line breaks, including actual hyphenations, the character representation of the text sequence should reflect any changes due to hyphenation (for example: "tugg{2010}" / "gumi", where / represents the line break).

NOTE – The notations {00AD} and {2010} indicate the inclusion of the corresponding code points: 00AD and 2010 into the code unit sequences. The curly brackets "{}" are not part of the code unit sequences.

F.1.2 Word boundary indicators

ZERO WIDTH SPACE (200B): This character behaves like a SPACE in that it indicates a word boundary, but unlike SPACE it has no presentational width. For example, this character could be used to indicate word boundaries in Thai, which does not use visible gaps to separate words.

WORD JOINER (2060) and **ZERO WIDTH NO-BREAK SPACE** (FEFF): These characters behave like a NO-BREAK SPACE in that they indicate the absence of word boundaries, but unlike NO-BREAK SPACE they have no presentational width. For example, these characters could be inserted after the fourth character in the text "base+delta" to indicate that there is to be no word break between the "e" and the "+".

NOTE – For additional usages of the ZERO WIDTH NO-BREAK SPACE for "signature", see Clause 10.

F.1.3 Cursive joiners

The characters described in this sub-clause are used to indicate whether or not the adjacent characters are joined together in rendering (cursive joiners).

ZERO WIDTH NON-JOINER (200C): This character indicates that the adjacent characters are not joined together in cursive connection even when they would normally join together as cursive letter forms. For example, ZERO WIDTH NON-JOINER between ARABIC LETTER NOON and ARABIC LETTER MEEM indicates that the characters are not rendered with the normal cursive connection.

MONGOLIAN VOWEL SEPARATOR (180E): This character acts in a similar fashion to the ZERO WIDTH NON-JOINER by suppressing the cursive connection between adjacent characters but only in the context of Mongolian vowels.

ZERO WIDTH JOINER (200D): This character indicates that the adjacent characters are represented with joining forms in cursive connection even when they would not normally join together as cursive letter forms. For example, in the sequence SPACE followed by ARABIC LETTER BEH followed by SPACE, ZERO WIDTH JOINER can be inserted between the first two characters to display the final form of the ARABIC LETTER BEH.

F.1.4 Format separators

The characters described in this sub-clause are used to indicate formatting boundaries between lines or paragraphs.

LINE SEPARATOR (2028): This character indicates where a new line starts; although the text continues to the next line, it does not start a new paragraph; e.g. no inter-paragraph indentation might be applied.

PARAGRAPH SEPARATOR (2029): This character indicates where a new paragraph starts; e.g. the text continues on the next line and inter-paragraph line spacing or paragraph indentation might be applied.

F.1.5 Bidirectional text formatting

The characters described in this sub-clause are used in formatting bidirectional text. If the specification of a subset includes these characters, then texts containing right-to-left characters are to be rendered with an implicit bidirectional algorithm.

An implicit algorithm uses the directional character properties to determine the correct display order of characters on a horizontal line of text.

The following three characters are format characters that act exactly like right-to-left or left-to-right characters in terms of affecting ordering (Bidirectional format marks). They have no visible graphic symbols, and they do not have any other semantic effect.

Their use can be more convenient than the explicit embeddings or overrides, since their scope is more local.

LEFT-TO-RIGHT MARK (200E): In bidirectional formatting, this character acts like a left-to-right character (such as LATIN SMALL LETTER A).

ARABIC LETTER MARK (061C): In bidirectional formatting, this character acts like a right-to-left Arabic character (such as ARABIC LETTER NOON).

RIGHT-TO-LEFT MARK (200F): In bidirectional formatting, this character acts like a generic right-to-left character (such as NKO LETTER A).

The following five format characters indicate that a piece of text is to be treated as embedded, and is to have a particular ordering attached to it (Bidirectional format embeddings). For example, an English quotation in the middle of an Arabic sentence can be marked as being an embedded left-to-right string. These format characters nest in blocks, with the embedding and override characters initiating (pushing) a block, and the pop character terminating (popping) a block.

The function of the embedding and override characters are very similar; the main difference is that the embedding characters specify the implicit direction of the text, while the override characters specify the explicit direction of the text. When text has an explicit direction, the normal directional character properties are ignored, and all of the text is assumed to have the ordering direction determined by the override character.

LEFT-TO-RIGHT EMBEDDING (202A): This character is used to indicate the start of a left-to-right implicit embedding.

RIGHT-TO-LEFT EMBEDDING (202B): This character is used to indicate the start of a right-to-left implicit embedding.

LEFT-TO-RIGHT OVERRIDE (202D): This character is used to indicate the start of a left-to-right explicit embedding.

RIGHT-TO-LEFT OVERRIDE (202E): This character is used to indicate the start of a right-to-left explicit embedding.

POP DIRECTIONAL FORMATTING (202C): This character is used to indicate the termination of an implicit or explicit directional embedding initiated by one of the four characters above.

The following four format characters, commonly called isolate characters, can be applied to a text segment to reduce its effect on the bidirectional ordering of its surroundings to that of a neutral character. This is in contrast to the existing embedding formatting characters (LEFT-TO-RIGHT EMBEDDING, RIGHT-TO-LEFT EMBEDDING, POP DIRECTIONAL FORMATTING) which have the effect of a strong character on their surroundings. Otherwise, isolate characters are similar to embedding characters: they declare a direction for the text inside it, and can be nested inside another isolate or embedding (and vice-versa).

LEFT-TO-RIGHT ISOLATE (2066): This character is used to indicate the start of a left-to-right isolate.

RIGHT-TO-LEFT ISOLATE (2067): This character is used to indicate the start of a right-to-left isolate.

FIRST STRONG ISOLATE (2068): This character is used to indicate the start of a first-strong isolate, i.e. one whose direction is determined by applying specific Unicode Bidi Algorithm (see Clause 3) paragraph level rules to the isolate's content as if it were a separate paragraph.

POP DIRECTIONAL ISOLATE (2069): This character is used to indicate the end of an isolate.

F.2 Script-specific format characters

F.2.1 Symmetric swapping format characters

The following two characters are used in conjunction with the class of left/right handed pairs of mirrored characters described in Clause 15. The following format characters indicate whether the interpretation of the term LEFT or RIGHT in the character names is OPENING or CLOSING respectively. The following characters do not nest.

The default state of interpretation may be set by a higher level protocol or standard, such as ISO/IEC 6429. In the absence of such a protocol, the default state is as established by **ACTIVATE SYMMETRIC SWAPPING**.

INHIBIT SYMMETRIC SWAPPING (206A): Between this character and the following **ACTIVATE SYMMETRIC SWAPPING** format character (if any), the mirrored characters described in Clause 15 are interpreted and rendered as LEFT and RIGHT, and the processing specified in that clause is not performed.

ACTIVATE SYMMETRIC SWAPPING (206B): Between this character and the following **INHIBIT SYMMETRIC SWAPPING** format character (if any), the mirrored characters described in Clause 15 are interpreted and rendered as OPENING and CLOSING characters as specified in that clause.

F.2.2 Character shaping selectors

The following two characters are used in conjunction with Arabic presentation forms. During the presentation process, certain characters may be joined together in cursive connection or ligatures. The following characters indicate that the character shape determination process used to achieve this presentation effect is either activated or inhibited. The following characters do not nest.

INHIBIT ARABIC FORM SHAPING (206C): Between this character and the following **ACTIVATE ARABIC FORM SHAPING** format character (if any), the character shaping determination process is inhibited. The stored Arabic presentation forms are presented without shape modification. This is the default state.

ACTIVATE ARABIC FORM SHAPING (206D): Between this character and the following **INHIBIT ARABIC FORM SHAPING** format character (if any), the stored Arabic presentation forms are presented with shape modification by means of the character shaping determination process.

NOTE – These characters have no effect on characters that are not presentation forms: in particular, Arabic nominal characters as from 0600 to 06FF are always subject to character shaping, and are unaffected by these formatting characters.

F.2.3 Numeric shape selectors

The following two characters allow the selection of the shapes in which the digits from 0030 to 0039 are rendered. The following characters do not nest.

NATIONAL DIGIT SHAPES (206E): Between this character and the following **NOMINAL DIGIT SHAPES** format character (if any), digits from 0030 to 0039 are rendered with the appropriate national digit shapes as specified by means of appropriate agreements. For example, they could be displayed with shapes such as the ARABIC-INDIC digits from 0660 to 0669.

NOMINAL DIGIT SHAPES (206F): Between this character and the following **NATIONAL DIGIT SHAPES** format character (if any), the digits from 0030 to 0039 are rendered with the shapes as those shown in the code charts for those digits. This is the default state.

F.3 Interlinear annotation characters

The following three characters are used to indicate that an identified character string (the annotation string) is regarded as providing an annotation for another identified character string (the base string).

INTERLINEAR ANNOTATION ANCHOR (FFF9): This character indicates the beginning of the base string.

INTERLINEAR ANNOTATION SEPARATOR (FFFA): This character indicates the end of the base string and the beginning of the annotation string.

INTERLINEAR ANNOTATION TERMINATOR (FFFB): This character indicates the end of the annotation string.

The relationship between the annotation string and the base string is defined by agreement between the user of the originating device and the user of the receiving device. For example, if the base string is rendered in a visible form the annotation string may be rendered on a different line from the base string, in a position close to the base string.

If the interlinear annotation characters are filtered out during processing, then all characters between the Interlinear Annotation Separator and the Interlinear Annotation Terminator should also be filtered out.

F.4 Subtending format characters

The following nine characters are used to subtend a sequence of subsequent characters:

0600	ARABIC NUMBER SIGN
0601	ARABIC SIGN SANAH
0602	ARABIC FOOTNOTE MARKER
0603	ARABIC SIGN SAFHA
0604	ARABIC SIGN SAMVAT

ISO/IEC 10646:2014 (E)

0605	ARABIC NUMBER MARK ABOVE
06DD	ARABIC END OF AYAH
070F	SYRIAC ABBREVIATION MARK
110BD	KAITHI NUMBER SIGN

The scope of these characters and more details about their usage can be found in the Unicode Standard (see Annex M for referencing information).

F.5 Shorthand format characters

The use of overlapping letters to indicate abbreviations and initialisms is found in many shorthand systems. The following two characters are used to control such overlap.

SHORTHAND FORMAT LETTER OVERLAP (1BCA0): This character indicates a single letter overlap, with the text continuing to flow as if that overlapping character did not exist.

SHORTHAND FORMAT CONTINUING OVERLAP (1BCA1): This character indicates a continuing overlap where the text flow proceeds from the overlapping character.

In some shorthand systems certain set of word endings are indicated by letters following not in the default direction of text flow - to the right, but above or below the word. The following two characters are used to control such behaviours.

SHORTHAND FORMAT DOWN STEP (1BCA2): This character indicates that a following character should be rendered below the previous character, with any subsequent joined characters proceeding relative to the lowered glyph. . At word boundaries, this causes the next word (or stenographic period) to be lowered.

SHORTHAND FORMAT UP STEP (1BCA3): This character indicates that a following word (or stenographic period) is to be raised.

F.6 Invisible mathematical operators

In mathematics, some operators and punctuation are often implied but not displayed. Special format control characters known as invisible operators can be used to make such operators explicit for use in machine interpretation of mathematical expressions.

FUNCTION APPLICATION (2061): This character indicates the application of a function.

INVISIBLE TIMES (2062): This character indicates a multiplication.

INVISIBLE SEPARATOR (2063): This character indicates that adjacent mathematical symbols form a list, e.g. when no visible COMMA is used between multiple indices.

INVISIBLE PLUS (2064): This character indicates an addition.

F.7 Western musical symbols

This international standard does not specify an encoding solution for musical scores or musical pitch. Solutions for these needs would require another description layer on top of the encoding definition of the characters specified in this standard. However, even without that additional layer, these characters can be used as simple musical reference symbols for general purposes in text descriptions of musical matters.

Extended beams are used frequently in music notation between groups of notes having short values. The format characters 1D173 MUSICAL SYMBOL BEGIN BEAM and 1D174 MUSICAL SYMBOL END BEAM can be used to indicate the extents of beam groupings. In some exceptional cases, beams are unclosed on one end. This can be indicated with a "null note" (MUSICAL SYMBOL NULL NOTEHEAD) character if no stem is to appear at the end of the beam.

Similarly, other format characters have been provided for other connecting structures. The characters

1D175 MUSICAL SYMBOL BEGIN TIE
 1D176 MUSICAL SYMBOL END TIE
 1D177 MUSICAL SYMBOL BEGIN SLUR
 1D178 MUSICAL SYMBOL END SLUR
 1D179 MUSICAL SYMBOL BEGIN PHRASE
 1D17A MUSICAL SYMBOL END PHRASE

indicate the extent of these features.

These pairs of characters modify the layout and grouping of notes and phrases in full music notation. When musical examples are written or rendered in plain text without special software, the start/end control characters may be rendered as brackets or left un-interpreted. More sophisticated in-line processes may interpret them, to the extent possible, in their actual control capacity, rendering ties, slurs, beams, and phrases as appropriate.

For maximum flexibility, the character set includes both pre-composed note values as well as primitives from which complete notes are constructed. Due to their ubiquity, the pre-composed versions are provided mainly for convenience.

Coding convenience notwithstanding, notes built up from alternative noteheads, stems and flags, and articulation symbols are necessary for complete implementations and complex scores. Examples of their use include American shape-note and modern percussion notations. For example,

MUSICAL SYMBOL SQUARE NOTEHEAD BLACK + MUSICAL SYMBOL COMBINING STEM

MUSICAL SYMBOL X NOTEHEAD + MUSICAL SYMBOL COMBINING STEM

Augmentation dots and articulation symbols may be appended to either the pre-composed or built-up notes.

In addition, augmentation dots and articulation symbols may be repeated as necessary to build a complete note symbol. For example,

MUSICAL SYMBOL EIGHTH NOTE + MUSICAL SYMBOL COMBINING AUGMENTATION DOT + MUSICAL SYMBOL COMBINING AUGMENTATION DOT + MUSICAL SYMBOL COMBINING ACCENT

F.8 Language tagging using Tag characters

F.8.1 General

The purpose of Tag characters is to associate a text attribute with a point or range of a text string. The value of a particular tag is not generally considered to be part of the content of the text. For example, tagging could be used to mark the language or the font applied to a portion of text. Outside of that usage, these characters are ignorable.

These tag characters can be used to spell out a character string in any ASCII-based tagging scheme that needs to be embedded into plain text. These characters can be easily identified by their code value and there is no overloading of usage for these tag characters. They can only express tag values and never textual content itself.

When characters are used within the context of a protocol or syntax containing explicit markup providing the same association, the Tag characters may be filtered out and ignored by these protocols.

For example, in SGML/XML context, an explicit language markup is specified. Therefore, the LANGUAGE TAG (E0001) and other tag characters should not be used to mark a language in that context. The Unicode Consortium and the W3C have co-written a technical report: Unicode in XML and other Markup Languages (UTR#20), available from the Unicode web site (<http://www.unicode.org/reports/>), which describes these issues in detail.

The TAGS block contains 97 dedicated tag characters consisting of a clone of the BASIC LATIN graphic characters (names formed by prefixing these BASIC LATIN names with the word 'TAG', code points from

E0020 to E007E), as well as a language tag identification character: LANGUAGE TAG (E0001) and a cancel tag character: CANCEL TAG (E007F).

The tag identification character is used as a mechanism for identifying tags of different types. This enables multiple types of tags to coexist amicably embedded in plain text and solves the problem of delimitation if a tag is concatenated directly onto another tag. Although only one type of tag is currently specified, namely the language tag, the encoding of other tag identification characters in the future would allow for distinct types to be used.

F.8.2 Syntax for embedding tag characters

In order to embed any ASCII-derived tag in plain text, the tag is simply spelled out with the tag characters, prefixed with the relevant tag identification character. The resultant string is embedded directly in the text.

No termination character is required for a tag. A tag terminates either when the first non Special Purpose Plane character is encountered, or when the next tag identification character is encountered.

Tag arguments can only be encoded using tag characters. No other characters are valid for expressing the tag arguments.

F.8.3 Tag scope and nesting

The value of a tag continues from the point the tag is embedded in text until

- either the end of the code unit sequence is reached,
- or the tag is explicitly cancelled by the CANCEL TAG character.

Tags of the same type cannot be nested. The appearance of a new embedded language tag, for example after text which was already language-tagged, simply changes the tagged value for subsequent text to that specified in the new tag.

F.8.4 Cancelling tag values

The CANCEL TAG character is provided to allow the specific canceling of a tag value. For example to cancel a language tag, the LANGUAGE TAG should precede the CANCEL TAG character.

The usage of the CANCEL TAG character without a prefixed tag identification character cancels any tag value that may be defined.

The main function of the character is to make possible such operations as blind concatenation of strings in a tagged context without the propagation of inappropriate tag values across the string boundaries.

F.8.5 Language tags

Language tags are of general interest and may have a high degree of interoperability for protocol usage. For example, to embed a language tag for Japanese, the tag characters would be used as follows:

E0001 E006A E0061

The first value is the coded value of the LANGUAGE TAG character, the second corresponds to the TAG LATIN SMALL LETTER J, and the third corresponds to the TAG LATIN SMALL LETTER A. The sequence 'ja' corresponds to the 2-letter code representing the Japanese language in ISO 639:1988.

Annex G
(informative)
Alphabetically sorted list of character names

The alphabetically sorted list of character names is provided in machine-readable format that is accessible as a link to this document. The content linked to is a plain text file, using ISO/IEC 646-IRV characters with CARRIAGE RETURN/ LINE FEED as end of line mark, that specifies, after a 4-lines header, all the character names from this International Standard except Hangul syllables and CJK ideographs (these are characters from blocks:

HANGUL SYLLABLES,
CJK UNIFIED IDEOGRAPHS,
CJK UNIFIED IDEOGRAPHS EXTENSION A,
CJK UNIFIED IDEOGRAPHS EXTENSION B,
CJK UNIFIED IDEOGRAPHS EXTENSION C,
CJK UNIFIED IDEOGRAPHS EXTENSION D,
CJK UNIFIED IDEOGRAPHS EXTENSION E,
CJK COMPATIBILITY IDEOGRAPHS, and
CJK COMPATIBILITY IDEOGRAPHS SUPPLEMENT).

Each line contains the following information organized in fields delimited by the TAB character:

- 1st field: UCS code point in the format (hhhh | hhhhh), 'h' being a hexadecimal unit,
- 2nd field: character name.

[Click on this highlighted text to access the reference file.](#)

NOTE 1 – The content is also available as a separate viewable file in the same file directory as this document. The file is named: "Allnames.txt".

NOTE 2 – The referenced files are only available to users who obtain their copy of the standard in a machine-readable format. However, the file format makes them printable.

Annex H
(informative)
The use of “signatures” to identify UCS

NOTE – Integrated in main body text, see Clause 10.

Annex I (informative) Ideographic description characters

I.1 General

An Ideographic Description Character (IDC) is a graphic character, which is used with a sequence of other graphic characters to form an Ideographic Description Sequence (IDS). Such a sequence may be used to describe an ideographic character which is not specified within this International Standard.

The IDS describes the ideograph in the abstract form. It is not interpreted as a composed character and does not imply any specific form of rendering.

NOTE – An IDS is not a character and therefore is not a member of the repertoire of this International Standard.

I.2 Syntax of an ideographic description sequence

An IDS consists of an IDC followed by a fixed number of Description Components (DC). A DC may be any one of the following:

- a coded ideograph
- a coded radical
- the character FF1F FULLWIDTH QUESTION MARK to represent an otherwise un-described DC
- a private use character (as long as the interchanging parties have agreed that the particular private use character represents a particular ideograph or component of an ideograph)
- another IDS

NOTE 1 – The above description implies that any IDS may be nested within another IDS.

Each IDC has four properties as summarized in table I.1 below;

- the number of DCs used in the IDS that commences with that IDC,
- the definition of its acronym,
- the syntax of the corresponding IDS,
- the relative positions of the DCs in the visual representation of the ideograph that is being described in its abstract form.

The syntax of the IDS introduced by each IDC is indicated in the “IDS Acronym and Syntax” column of the table by the abbreviated name of the IDC (e.g. IDC-LTR) followed by the corresponding number of DCs, i.e. (D₁ D₂) or (D₁ D₂ D₃).

I.3 Individual definitions of the ideographic description characters

IDEOGRAPHIC DESCRIPTION CHARACTER LEFT TO RIGHT (2FF0): The IDS introduced by this character describes the abstract form of the ideograph with D₁ on the left and D₂ on the right.

IDEOGRAPHIC DESCRIPTION CHARACTER ABOVE TO BELOW (2FF1): The IDS introduced by this character describes the abstract form of the ideograph with D₁ above D₂.

IDEOGRAPHIC DESCRIPTION CHARACTER LEFT TO MIDDLE AND RIGHT (2FF2): The IDS introduced by this character describes the abstract form of the ideograph with D₁ on the left of D₂, and D₂ on the left of D₃.

IDEOGRAPHIC DESCRIPTION CHARACTER ABOVE TO MIDDLE AND BELOW (2FF3): The IDS introduced by this character describes the abstract form of the ideograph with D₁ above D₂, and D₂ above D₃.

IDEOGRAPHIC DESCRIPTION CHARACTER FULL SURROUND (2FF4): The IDS introduced by this character describes the abstract form of the ideograph with D_1 surrounding D_2 .

IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM ABOVE (2FF5): The IDS introduced by this character describes the abstract form of the ideograph with D_1 above D_2 , and surrounding D_2 on both sides.

IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM BELOW (2FF6): The IDS introduced by this character describes the abstract form of the ideograph with D_1 below D_2 , and surrounding D_2 on both sides.

IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LEFT (2FF7): The IDS introduced by this character describes the abstract form of the ideograph with D_1 on the left of D_2 , and surrounding D_2 above and below.

IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM UPPER LEFT (2FF8): The IDS introduced by this character describes the abstract form of the ideograph with D_1 at the top left corner of D_2 , and partly surrounding D_2 above and to the left.

IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM UPPER RIGHT (2FF9): The IDS introduced by this character describes the abstract form of the ideograph with D_1 at the top right corner of D_2 , and partly surrounding D_2 above and to the right.

IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LOWER LEFT (2FFA): The IDS introduced by this character describes the abstract form of the ideograph with D_1 at the bottom left corner of D_2 , and partly surrounding D_2 below and to the left.

IDEOGRAPHIC DESCRIPTION CHARACTER OVERLAID (2FFB): The IDS introduced by this character describes the abstract form of the ideograph with D_1 and D_2 overlaying each other.

Table I.1: Properties of ideographic description characters

Character Name: IDEOGRAPHIC DESCRIPTION CHARACTER ...	no. of DCs	IDS Acronym and Syntax	Relative posi- tions of DCs	Example of IDS	IDS example represents:
LEFT TO RIGHT	2	IDC-LTR D ₁ D ₂			𠂇母
ABOVE TO BELOW	2	IDC-ATB D ₁ D ₂			𠂇天
LEFT TO MIDDLE AND RIGHT	3	IDC-LMR D ₁ D ₂ D ₃			𠂇言子
ABOVE TO MIDDLE AND BELOW	3	IDC-AMB D ₁ D ₂ D ₃			𠂇从日
FULL SURROUND	2	IDC-FSD D ₁ D ₂			𠂇莫
SURROUND FROM ABOVE	2	IDC-SAV D ₁ D ₂			𠂇門卜
SURROUND FROM BELOW	2	IDC-SBL D ₁ D ₂			𠂇山士
SURROUND FROM LEFT	2	IDC-SLT D ₁ D ₂			𠂇匚虎
SURROUND FROM UPPER LEFT	2	IDC-SUL D ₁ D ₂			𠂇广舞
SURROUND FROM UPPER RIGHT	2	IDC-SUR D ₁ D ₂			𠂇勺去
SURROUND FROM LOWER LEFT	2	IDC-SLL D ₁ D ₂			𠂇辶交
OVERLAID	2	IDC-OVL D ₁ D ₂			𠂇从工
NOTE – In IDC-OVL, D ₁ and D ₂ overlap each other. This diagram does not imply that D ₁ is on the top left corner and D ₂ is on the bottom right corner.					

Annex J
(informative)

Recommendation for combined receiving/originating devices with internal storage

Annex J is applicable to a widely-used class of devices that can store received code unit sequences for subsequent retransmission.

This recommendation is intended to ensure that loss of information is minimized between the receipt of a code unit sequence and its retransmission.

A device of this class includes a receiving device component and an originating device component as in 2.3, and can also store received code unit sequences for retransmission, with or without modification by the actions of the user on the corresponding characters represented within it. Within this class of device, two distinct types are identified here, as follows.

- 1) **Receiving device with full retransmission capability.** The originating device component will retransmit the coded representations of any received characters, including those that are outside the identified subset of the receiving device component, without change to their coded representation, unless modified by the user.
- 2) **Receiving device with subset retransmission capability.** The originating device component can retransmit only the coded representations of the characters of the subset adopted by the receiving device component.

Annex K
(informative)
Notations of octet value representations

Representation of octet values in this International Standard except in Clause 12 is different from other character coding standards such as ISO/IEC 2022, ISO/IEC 6429 and ISO 8859. Annex K clarifies the relationship between the two notations.

In this International Standard, the notation used to express an octet value is z , where z is a hexadecimal number in the range 00 to FF. For example, the character ESCAPE (ESC) of ISO/IEC 2022 is represented in this International Standard by 1B.

In other character coding standards, the notation used to express an octet value is x/y , where x and y are two decimal numbers in the range 00 to 15. The correspondence between the notations of the form x/y and the octet value is as follows.

- x is the number represented by bit 8, bit 7, bit 6 and bit 5 where these bits are given the weights 8, 4, 2 and 1 respectively;
- y is the number represented by bit 4, bit 3, bit 2 and bit 1 where these bits are given the weights 8, 4, 2 and 1 respectively.

For example, the character ESC of ISO/IEC 2022 is represented by 01/11.

Thus ISO/IEC 2022 (and other character coding standards) octet value notation can be converted to octet value notation used by this International Standard by converting the value of x and y to hexadecimal notation. For example; 04/15 is equivalent to 4F.

Annex L (informative) Character naming guidelines

The Clause 24 of this standard specifies rules for name formation and name uniqueness. These rules are used in other information technology coded character set standards such as ISO/IEC 646, ISO/IEC 6937, ISO/IEC 8859, and ISO/IEC 10367. Annex L provides additional guidelines for the creation of these entity names.

These guidelines do not apply to the names of CJK Ideographs and Hangul syllables which are formed using rules specified in 24.6 and 24.7 respectively.

Guideline 1

The name of an entity wherever possible denotes its customary meaning (for example, the character name: PLUS SIGN or the block name: BENGALI).

Some entities, such as characters, may have a name describing shapes, not usage, (for example, the character name: UPWARDS ARROW).

The name on an entity is not intended to identify its properties or attributes, or to provide information on its linguistic characteristics, except as defined in guideline 4 below.

Guideline 2

An acronym consists of Latin capital letters A to Z and digits and is associated with a name.

Acronyms may be used in entity names where usage already exists and clarity requires it. For example, the names of control functions are coupled with an acronym.

EXAMPLES

<u>Name:</u>	<u>Acronym</u>
LOCKING-SHIFT TWO RIGHT	LS2R
SOFT HYPHEN	SHY
INTERNATIONAL PHONETIC ALPHABET	IPA

NOTE – In ISO/IEC 6429, the names of the modes have also been presented in the same way as control functions.

Guideline 3

Character names and Named UCS Sequence Identifiers (NUSI) only include digits 0 to 9 if spelling out the name of the corresponding digits(s) would be inappropriate.

NOTE – As an example the name of the character at the code point value 201A is SINGLE LOW-9 QUOTATION MARK; the symbol for the digit 9 is included in this name to illustrate the shape of the character, and has no numerical significance.

Guideline 4

Character names and NUSIs are constructed from an appropriate set of the applicable terms of the following grid and ordered in the sequence of this grid. Exceptions are specified in guidelines 9 to 11. The words WITH and AND may be included for additional clarity when needed.

1	Script	5	Attribute
2	Case	6	Designation
3	Type	7	Mark(s)
4	Language	8	Qualifier

EXAMPLES OF SUCH TERMS

Script	Latin, Cyrillic, Arabic
Case	capital, small
Type	letter, ligature, digit
Language	Ukrainian
Attribute	final, sharp, subscript, vulgar
Designation	customary name, name of letter
Mark	acute, ogonek, ring above, diaeresis
Qualifier	sign, symbol

EXAMPLES OF NAMES

LATIN CAPITAL LETTER A WITH ACUTE

1 2 3 6 7

DIGIT FIVE

3 6

LEFT CURLY BRACKET

5 5 6

NOTE – A ligature is a graphic symbol in which two or more other graphic symbols are imaged as a single graphic symbol.

For character names, where a character comprises a base letter with multiple marks, the sequence of those in the name is the order in which the marks are positioned relative to the base letter. The sequence may start with the marks above the letters taken in upwards sequence, and follow with the marks below the letters taken in downwards sequence, or the reverse (below/above).

For NUSIs, where the sequence comprises a base letter with multiple marks, the name describes the individual characters in the sequence in which they are encoded in the sequence.

EXAMPLES

Ō LATIN CAPITAL LETTER O WITH CIRCUMFLEX AND DOT BELOW

Ç LATIN CAPITAL LETTER C WITH CEDILLA AND ACUTE

Ų LATIN CAPITAL LETTER U WITH OGONEK AND ACUTE

Guideline 5

The letters of the Latin script are represented within their name by their basic graphic symbols (A, B, C, etc.). The letters of all other scripts are represented by their transcription in the language of the first published International Standard.

EXAMPLES

K LATIN CAPITAL LETTER K
Ю CYRILLIC CAPITAL LETTER YU

Guideline 6

In principle when a character of a given script is used in more than one language, no language name is specified. Exceptions are tolerated where an ambiguity would otherwise result.

EXAMPLES

И CYRILLIC CAPITAL LETTER I
І CYRILLIC CAPITAL LETTER BYELORUSSIAN-UKRAINIAN I

Guideline 7

Letters that are elements of more than one script are considered different even if their shape is the same; they have different names.

EXAMPLES

A LATIN CAPITAL LETTER A
Α GREEK CAPITAL LETTER ALPHA
А CYRILLIC CAPITAL LETTER A

Guideline 8

Where possible, NUSIs are constructed by appending the names of the constituent elements together while eliding duplicate elements. Should this process result in a name that already exists, the name is modified suitably to guarantee uniqueness among character names and NUSIs. The words WITH and AND may be included for additional clarity when needed.

Guideline 9

A character of one script used in isolation in another script, for example as a graphic symbol in relation with physical units of dimension, is considered as a character different from the character of its native script.

EXAMPLE

μ MICRO SIGN

Guideline 10

A number of characters have a traditional name consisting of one or two words. It is not intended to change this usage.

EXAMPLES

'	APOSTROPHE
:	COLON
@	COMMERCIAL AT
_	LOW LINE
~	TILDE

Guideline 11

In some cases, characters of a given script, often punctuation marks, are used in another script for a different usage. In these cases the customary name reflecting the most general use is given to the character. The customary name may be annotated in the list of characters with the script name followed by the transliterated name in that script.

EXAMPLE

203F ᳵ UNDERTIE
= Greek Enotikon

Annex M (informative) Sources of characters

Several sources and contributions were used for constructing this coded character set. The sources are grouped by their categories. National and international standards are listed first for each category, followed by relevant publications references.

General

ISO international register of character sets to be used with escape sequences. (registration procedure ISO 2375:1985) .

ISO 8879:1986, *Information processing - Text and office systems - Standard Generalized Markup Language (SGML)*.

ISO/IEC TR 15285:1998, *Information technology - An operational model for characters and glyphs*.

JIS X 0201-1976 Japanese Standards Association. *Jouhou koukan you fugou (Code for Information Interchange)*.

Allworth, Edward. *Nationalities of the Soviet East: Publications and Writing Systems*. New York, London, Columbia University Press, 1971. ISBN 0-231-03274-9.

Barry, Randall K. 1997. *ALA-LC romanization tables: transliteration schemes for non-Roman scripts*. Washington, DC: Library of Congress Cataloging Distribution Service. ISBN 0-8444-0940-5

Daniels, Peter T., and William Bright, eds. 1996. *The world's writing systems*. New York; Oxford: Oxford University Press. ISBN 0-19-507993-0

Diringer, David. 1996. *The alphabet: a key to the history of mankind*. New Delhi: Munshiram Manoharlal. ISBN 81-215-0780-0

Faulmann, Carl. 1990 (1880). *Das Buch der Schrift*. Frankfurt am Main: Eichborn. ISBN 3-8218-1720-8

Haarmann, Harald. 1990. *Universalgeschichte der Schrift*. Frankfurt/Main; New York: Campus. ISBN 3-593-34346-0

Imprimerie Nationale. 1990. *Les caractères de l'Imprimerie nationale*. Paris: Imprimerie nationale Éditions. ISBN 2-11-081085-8

Jensen, Hans. 1969. *Die Schrift in Vergangenheit und Gegenwart*. 3., neu bearbeitete und erweiterte Auflage. Berlin: VEB Deutscher Verlag der Wissenschaften.

Knuth, Donald E. *The TeXbook*. – 19th. printing, rev, – Reading, MA : Addison-Wesley, 1990.

Nakanishi, Akira. 1990. *Writing systems of the world: alphabets, syllabaries, pictograms*. Rutland, VT: Charles E. Tuttle. ISBN 0-8048-1654-9

Shepherd, Walter. *Shepherd's glossary of graphic signs and symbols*. Compiled and classified for ready reference. – New York : Dover Publications, [1971].

The Unicode Consortium *The Unicode Standard. Worldwide Character Encoding Version 1.0, Volume One*. – Reading, MA : Addison-Wesley, 1991.

The Unicode Consortium *The Unicode Standard, Version 2.0*. Reading, MA: Addison-Wesley, 1996. ISBN 0-201-48345-9

ISO/IEC 10646:2014 (E)

The Unicode Consortium *The Unicode Standard, Version 3.0*. Reading, MA: Addison-Wesley Developer's Press, 2000. ISBN 0-201-61633-5

The Unicode Consortium *The Unicode standard, Version 4.0*. Reading, MA: Addison-Wesley Developer's Press, 2003. ISBN 0-321-18578-1

The Unicode Consortium *The Unicode Standard, Version 5.0*. Reading, MA: Addison-Wesley Developer's Press, 2007. ISBN 0-321-48091-0

The Unicode Consortium. *The Unicode Standard, Version 6.3*. Mountain View, CA: The Unicode Consortium, 2013. ISBN 978-1-936213-08-5 <http://www.unicode.org/versions/Unicode6.3.0/>

The Unicode Consortium. *The Unicode Standard, Version 7.0.0*, Mountain View, CA: The Unicode Consortium, 2014. ISBN 978-1-936213-09-2 <http://www.unicode.org/versions/Unicode7.0.0/>

Alchemical Symbols

Berthelot, Marcelin. *Collection des anciens alchimistes grecs*. 3 vols. Paris: G. Steinheil, 1888.

Berthelot, Marcelin. *La chimie au moyen âge*. 3 vols. Osnabrück: O. Zeller, 1967.

Lüdy-Tenger, Fritz. *Alchemistische und chemische Zeichen*. Würzburg: JAL-reprint, 1973.

Schneider, Wolfgang. *Lexicon alchemistisch-pharmazeutischer Symbole*. Weinheim/Bergstr.: Verlag Chemie, 1962.

Arabic

ISO 233:1984, *Documentation - Transliteration of Arabic characters into Latin characters*.

ISO/IEC 8859, *Information technology - 8-bit single-byte coded graphic character sets Part 6: Latin/Arabic alphabet* (1999)

ISO 9036:1987, *Information processing - Arabic 7-bit coded character set for information interchange*.

ASMO 449-1982 Arab Organization for Standardization and Metrology. *Data processing - 7-bit coded character set for information interchange*.

Avestan

Geldner, Karl F. *Avesta: The Sacred Books of the Parsis*. Stuttgart: W. Kohlhammer, 1880. Reprinted, with an introduction in Persian by Dr. Jaleh Amouzgar. Tehran: Asatir, 2003. ISBN 964-331-126-0.

Hoffmann, Karl, and B. Forssman. *Avestische Laut- und Flexionslehre*. Innsbruck: Innsbrucker Beiträge zur Sprachwissenschaft, 1996. ISBN 3851246527.

Oryan, Said. *Pahlavi-Pazand Glossary: Farhang \ Pahlavi*. Tehran: Research Institute for Islamic Culture and Art, 1999 (1377 AP). (Language and Literature, 4). ISBN 964-471-414-8.

Reichelt, Hans. *Avesta Reader: An Approach to the Zoroaster's Gathas and New Avestan Texts*. Translated and annotated with Persian translation of hymns and texts by Jalil Doostkhah. Tehran: Qoqnoos Publishing, 2004 (1383 AP). ISBN 964-311-473-2.

Balinese

Medra, Nengah. *Pedoman Pasang Aksara Bali*. Denpasar: Dinas Kebudayaan Propinsi Bali, 2003.

Menaka, Made. *Kamus Kawi Bali* / oleh, made Menaka. Singaraja: Yayasan Kawi Sastra Mandala, 1990.

Simpen, I Wayan. *Pasang Aksara Bali*. Denpasar: Upada Sastra, 1992.

Bamum

Dugast, J., and M. D. W. Jeffreys. *L'écriture des bamum: sa naissance, son évolution, sa valeur phonétique, son utilisation*. Mémoires de l'Institut Français d'Afrique Noire, Centre du Cameroun, 1950.

Nchare, Oumarou. *The Writing of King Njoya: Genesis, Evolution, Use*. Foumban: Palais des Rois Bamoun, Maison de la Culture, [s.d.].

Schmitt, Alfred. *Die Bamum-Schrift*. Band I: Text. Wiesbaden: Harrassowitz, 1963.

Batak

Kozok, Uli. *Warisan leluhur: sastra lama dan aksara Batak*. Jakarta: École française d'Extrême Orient, 1999. ISBN 979-9023-33-5.

Meerwaldt, J H. *Handleiding tot de beoefening der Bataksche taal*. Leiden: E.J. Brill, 1904.

Tuuk, Herman Neubronner van der. *A Grammar of Toba Batak*. Translated by Jeune Scott-Kemball, edited by Andries Teeuw and R. Roolvink. The Hague: Nijhoff, 1971.

Brahmi

Baums, Stefan. "Towards a Computer Encoding for Brāhmī." In *Script and Image: Papers on Art and Epigraphy*, edited by Adalbert J. Gail, Gerd J. R. Mevissen and Richard Salomon, vol. 11.1, 111–143. Delhi: Motilal Banarsidass Publishers, 2006.

Bühler, G. "The Bhattiprolu Inscriptions." In *Epigraphia Indica: A Collection of Inscriptions Supplementary to the Corpus Inscriptionum Indicarum of the Archaeological Survey*, vol. 2, 323–329. Calcutta: Epigraphia Indica, 1894.

Dani, Ahmad Hasan. *Indian Palaeography*. 2nd edition. New Delhi: Munshiram Manoharlal Publishers, 1986.

Mahadevan, Iravatham. *Early Tamil Epigraphy: From the Earliest Times to the Sixth Century A.D.* Chennai, India: Cre-A, 2003. (Harvard Oriental Series, vol. 62.)

Braille

ISO 11548-1:2001. *Communication aids for blind persons – identifiers, names and assignation to coded character sets for 8-dot Braille characters – Part 1: General guidelines for Braille identifiers and shift marks*.

Canadian Aboriginal Syllabics

Canadian Aboriginal Syllabic Encoding Committee. *Repertoire of Unified Canadian Aboriginal Syllabics Proposed for Inclusion into ISO/IEC 10646: International Standard Universal Multi-Octet Coded Character Set*. [Canada]: CASEC [1994]

Carian

Adiego, Ignacio-Javier. *The Carian Language*. Leiden; Boston: Brill, 2007.

Melchert, H. Craig. "Carian." In *The Cambridge Encyclopedia of the World's Ancient Languages*, edited by Roger Woodard, 609–613. Cambridge: Cambridge University Press, 2004. ISBN-13: 978-0521562560.

Chakma

Cāṇmā, Cirajyoti and Maṅgal Cāṅgmā. *Cāṇmār āg p u d h i* = Chakma primer. Rāṇamāṭi: Cāṇmābhāṣā Prakāśanā Pariṣad. 1982.

ISO/IEC 10646:2014 (E)

Khisa, Bhagadatta. *Cāṃmā pattham pāt* = Chakma primer. Rāṇamāṭi: Tribal Cultural Institute, 2001.

Cham

Aymonier, Étienne, and Antoine Cabaton. *Dictionnaire Čam-Français*. Paris, 1906.

Bùi Khánh Thế. *Từ điển Chăm-Việt: Inālang cam-biet đām*. [Hồ Chí Minh]: Nhà xuất bản Khoa Học Xã Hội, 1995.

Kōno Rokurō, Chino Eiichi, and Nishida Tatsuo. *The Sanseido Encyclopaedia of Linguistics*. Volume 7: *Scripts and Writing Systems of the World* [Gengogaku dai ziten (bekkan) sekai mozi ziten]. Tokyo: Sanseido Press, 2001. ISBN 4-385-15177-6.

Cherokee

Alexander, J. T. *A Dictionary of the Cherokee Indian Language*. [Sperry, Oklahoma?]: Published by the author, 1971.

Holmes, Ruth Bradley. *Beginning Cherokee*, by Ruth Bradley Holmes and Betty Sharp Smith. 2nd ed. Norman: University of Oklahoma Press, 1977. ISBN 0-8061-1464-9.

New Echota Letters: Contributions of Samuel A. Worcester to the Cherokee Phoenix, edited by Jack Frederick Kilpatrick and Anna Gritts Kilpatrick. Dallas: Southern Methodist University Press, [s.d.]. Includes reprint of an article by S. A. Worcester, which appeared in the *Cherokee Phoenix*, Feb. 21, 1828.

CJK Ideographs

GB2312-80 *Code of Chinese Graphic Character Set for Information Interchange: Jishu Biaozhun Chubanshe* (Technical Standards Publishing).

GBK (*Guo Biao Kuo*) *Han character internal code extension specification: Jishu Biaozhun Chubanshe* (Technical Standards Publishing, Beijing)

JIS X 0208-1990 Japanese Standards Association. *Jouhou koukan you kanji fugoukei (Code of the Japanese Graphic Character Set for Information Interchange)*.

JIS X 0212-1990 Japanese Standards Association. *Jouhou koukan you kanji fugou-hojo kanji (Code of the supplementary Japanese graphic character set for information interchange)*.

JIS X 0213:2000, Japanese Standards Association. *7-bit and 8-bit double byte coded extended KANJI sets for information interchange, 2000-01-20*.

KS X 1001:2004 (formerly KS C 56 01-1992) Korean Industrial Standards Association. *Code for Information Interchange (Hangeul and Hanja)* (Jeongbo gyohwanyong buhogye).

ANSI Z39.64-1989 American National Standards Institute. *East Asian character code for bibliographic use*.

Mandarin Promotion Council, Ministry of Education, Taiwan. *Shiangtu yuyan biauyin fuhau shoutse (The Handbook of Taiwan Languages Phonetic Alphabet)*. 1999.

Shinmura, Izuru. *Kojien – Dai 4-han*. – Tokyo : Iwanami Shoten, Heisei 3 [1991].

NOTE – For additional sources of the CJK unified ideographs in this International Standard refer to Clause 23.

Coptic

Browne, Gerald M. *Old Nubian Grammar*. München: Lincom Europa, 2002. (Languages of the world: Materials, 330). ISBN 3-89586-893-0 (pbk.).

Kasser, Rodolphe. "La 'Genève 1986': une nouvelle série de caractères typographiques coptes, proto-coptes et vieux-coptes créée à Genève." *Bulletin de la Société d'égyptologie de Genève*, 12 (1988): 59–60. ISSN 0255-6286.

Kasser, Rodolphe. "A standard system of Sigla for referring to the dialects of Coptic." *Journal of Coptic Studies*, 1 (1990): 141–151. ISSN 1016-5584.

Cypriot

See Linear B and Cypriot

Cyrillic

ISO/IEC 8859, *Information technology - 8-bit single-byte coded graphic character sets*
Part 5: *Latin/Cyrillic alphabet* (1999)

ISO 5427:1984, *Extension of the Cyrillic alphabet coded character set for bibliographic information interchange*.

ISO 10754:1984, *Information and documentation – Extension of the Cyrillic alphabet coded character set for non-Slavic languages for bibliographic information interchange*.

Deseret

Encyclopedia of Mormonism, entry for "Deseret Alphabet." New York: Macmillan, 1992. ISBN 0-02-904040-X.

Ivins, Stanley S. "The Deseret Alphabet" *Utah Humanities Review* 1 (1947): 223-39.

Monson, Samuel C. *Representative American Phonetic Alphabets*. New York: 1954. Ph.D. dissertation—Columbia University.

Egyptian Hieroglyphs

Allen, James P. Middle *Egyptian: An Introduction to the Language and Culture of Hieroglyphs*. Cambridge: Cambridge University Press, 1999. ISBN 0-521-77483-7.

Gardiner, Alan H. *Catalogue of the Egyptian Hieroglyphic Printing Type, from Matrices Owned and Controlled by Dr. Alan H. Gardiner, in Two Sizes, 18 Point, 12 Point with Intermediate Forms*. Oxford: Oxford University Press, 1928.

Gardiner, Alan H. "Additions to the New Hieroglyphic Fount (1928)." *The Journal of Egyptian Archaeology*, 15 (1929): 95. ISSN 0307 5133.

Gardiner, Alan H. "Additions to the New Hieroglyphic Fount (1931)." *The Journal of Egyptian Archaeology*, 17 (1931): 245–247. ISSN 0307 5133.

Gardiner, Alan H. *Supplement to the Catalogue of the Egyptian Hieroglyphic Printing Type, Showing Acquisitions to December 1953*. Oxford: Oxford University Press, 1953.

Gardiner, Alan H. *Egyptian Grammar: Being an Introduction to the Study of Hieroglyphs*. 3rd edition. London: Oxford University Press, 1957. ISBN 0-900416-35-1.

Ethiopic

Armbruster, Carl Hubert. *Initia Amharica: an Introduction to Spoken Amharic*. Cambridge, Cambridge University Press, 1908-20.

Launhardt, Johannes. *Guide to Learning the Oromo (Galla) Language*. Addis Ababa, Launhardt [1973?]

ISO/IEC 10646:2014 (E)

Leslau, Wolf. *Amharic Textbook*. Weisbaden, Harrassowitz; Berkeley, University of California Press, 1968.

Glagolitic

ISO 6861, *Information and documentation - Glagolitic alphabet coded character set for bibliographic information interchange*.

Glagolitica: zum Ursprung der slavischen Schriftkultur, herausgegeben von Heinz Miklas, unter der Mitarbeit von Sylvia Richter und Velizar Sadovski. Wien: Verlag der Österreichischen Akademie der Wissenschaften, 2000. (*Schriften der Balkan-Kommission, Philologische Abteilung*, 41). ISBN 3-7001-2895-9.

Khaburgaev, Georgii Aleksandrovich. *Staroslavianskii iazyk*. Izd. 2-e, perer. i dop. Moskva: Prosveshchenie, 1986.

Žubrinic, Darko. *Hrvatska glagoljica: biti pismen—biti svoj*. Zagreb: Hrvatsko književno društvo sv. Jeronima (sv. Cirila i Metoda): Element, 1996. ISBN 953-6111-35-7.

Gothic

Ebbinghaus, Ernst. "The Gothic Alphabet." In *The World's Writing Systems*, edited by Peter T. Daniels and William Bright. New York: Oxford University Press, 1996. ISBN 0-19-507993-0.

Fairbanks, Sydney, and F. P. Magoun Jr. 1940. 'On writing and printing Gothic', in *Speculum* 15:313-16.

Greek

ISO 5428:1984, *Greek alphabet coded character set for bibliographic information interchange*.

ISO/IEC 8859, *Information technology - 8-bit single-byte coded graphic character sets*
Part 7: Latin/Greek alphabet (1999)

Greek Editorial Marks

Austin, Colin. *Comicorum Graecorum Fragmenta in Papyris Reperta*, ed. Colinus Austin. Berolini [Berlin], Novi Eboraci [New York]: de Gruyter, 1973, p. 29. ISBN 3110024012.

Homer. *Iliad. Homeri Ilias*, edidit Thomas W. Allen. 3 vols. Oxonii [Oxford]: e typographeo Clarendoniano [Clarendon Press], 1931, vol. 2: pp. 39, 234.

The Oxyrhynchus Papyri, Part XV, edited with translations and notes by Bernard P. Grenfell and Arthur S. Hunt. London: Egypt Exploration Society, 1921, p. 56. (*Egypt Exploration Society, Graeco-Roman Memoirs*, 18).

Hebrew

ISO/IEC 8859, *Information technology - 8-bit single-byte coded graphic character sets*
Part 8: Latin/Hebrew alphabet (1999)

ISO 8957:1996, *Information and documentation - Hebrew alphabet coded character sets for bibliographic information interchange*.

SI 1311.1 – 1996: Standards Institution of Israel. Information technology. *ISO 8 bit coded character set with Hebrew points*.

SI 1311.2 – 1996, The Standards Institution of Israel. Information Technology. *ISO 8-bit coded character set for information interchange with Hebrew points and cantillation marks*.

Imperial Aramaic

Driver, G. R. *Semitic Writing from Pictograph to Alphabet*. 3rd ed. by S. A. Hopkins. London: Oxford University Press for the British Academy, 1976. ISBN 9780197259177.

Lidzbarski, Mark. *Handbuch der nordsemitischen Epigraphik nebst ausgewählten Inschriften*. Hildesheim: Georg Olms Verlagsbuchhandlung, 1962. Reprint of 1898 edition.

Naveh, Joseph. *Early History of the Alphabet: An Introduction to West Semitic Epigraphy and Palaeography*. Jerusalem: Magnes Press, the Hebrew University, 1987. ISBN 965-223-436-2.

Porten, Bezalel, and Ada Yardeni. *Textbook of Aramaic Documents from Ancient Egypt*. 4 vols. Jerusalem: Hebrew University, 1986–1999. ISBN 9652220752 (v. 1), 96 53500031 (v. 2), 9653500147 (v. 3), 9653500899 (v. 4).

Rosenthal, Franz. *A Grammar of Biblical Aramaic*. 7th rev. ed. Wiesbaden: Harrassowitz, 2006. ISBN 3-447-05251-1.

Inscriptional Parthian and Inscriptional Pahlavi

Akbarzādeh, Dāriyūš. *Katibe-hā-ye Pahlavi-ye Aškāni (Pārti) = Parthian Inscriptions*. Vol. 2. Tehran: Pazineh Press, 2002 (1381 AP). ISBN 964-5722-74-8.

Akbarzādeh, Dāriyūš. *Katibe-hā-ye Pahlavi: sang-negāre, sekke, mohr, asar-e mohr, zarfnebešte = Pahlavi Inscriptions: Inscriptions, Coins, Seals, Sealing Impression*. Vol. I. Tehran: Pazineh Press, 2003 (1382 AP). ISBN 964-5722-44-6.

Nyberg, Henrik Samuel. *A Manual of Pahlavi*. 2 vols. Wiesbaden: Harrassowitz, 1964–1974. ISBN 9783447015806 (vol. 2). Reprinted: Tehran: Asatir, 2003. ISBN 964-331-132-5, 964-331-131-7.

Oryan, Saeed. *Rahnāmā-ye katibe-hā-ye Irāni-ye miyāne Pahlavi-Pārti = Manual of Middle Iranian Inscriptions (Parthian-Pahlavi)*. Tehran: Iranian Cultural Heritage Organization, 2003 (1382 AP). ISBN 964-7483-71-6.

Rezai Baghbidi, Hassan. *Dastur-e Zabān-e Pārti (Pahlavi-e Aškāni) = A Grammar of Parthian (Arsacid Pahlavi)*. Iranian Academy of Persian Language and Literature, 2002 (1381 AP). ISBN 964-7531-05-2.

Indian scripts

IS 13194:1991 Bureau of Indian Standards *Indian script code for information interchange - ISCII*

LTD 37(1610)-1988 *Indian standard code for information interchange*.

International Phonetic Alphabet

Esling, John. *Computer coding of the IPA: supplementary report*. Journal of the International Phonetic Association, 20:1 (1990), p. 22-26.

International Phonetic Association. The IPA 1989 Kiel Convention Workgroup 9 report: *Computer Coding of IPA Symbols and Computer Representation of Individual Languages*. Journal of the International Phon. Assoc., 19:2 (1989), p. 81-82.

International Phonetic Association. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press, 1999. ISBN 0-521-65236-7; 0-521-63751-1 (pbk.).

International Phonetic Association. <http://www2.arts.gla.ac.uk/IPA/ipa.html>.

Journal of the International Phonetic Association, 24:2 (1994), 95–98, and 25:1 (1995), 21.

Pullum, Geoffrey K. *Remarks on the 1989 revision of the International Phonetic Alphabet*. Journal of the International Phonetic Association, 20:1 (1990), p. 33-40.

ISO/IEC 10646:2014 (E)

Pullum, Geoffrey K., and William A. Ladusaw. *Phonetic Symbol Guide*. 2nd ed. Chicago: University of Chicago Press, 1996. ISBN 0-226-68535-7; 0-226-68536-5 (pbk.).

Wells, John Christopher. *Accents of English*. Cambridge, New York: Cambridge University Press, 1982. Vol. 1: *Introduction*. ISBN 0-521-22919-7; ISBN 0-521-29719-2 (pbk.); vol. 2: *The British Isles*. ISBN 0-521-24224-X, ISBN 0-521-28540-2 (pbk.); vol. 3: *Beyond the British Isles*. ISBN 0-521-24225-8, ISBN 0-521-28541-0 (pbk.).

Javanese

Bohatta, Hanns. *Praktische Grammatik der javanischen Sprache, mit Lesestücken, einem javanisch-deutschen und deutsch-javanischen Wörterbuch*. Wien, Pest, Leipzig: Hartleben, [1892]. (Kunst der Polyglottie, 39).

Rochadi GK, R. H., and R. L. Sadeli Erawan BK. *Cacarakan aksara Sunda*. Bandung: Harisma, 1984.

Roorda, T. *Javaansche grammatica, benevens een leesboek tot oefening in de javaansche taal*. Amsterdam: Johannes Müller, 1855.

Walbeehm, A. H. J. G. *Javaansche spraakkunst: schrift, uitspraak, taalsoorten en woordafleiding*. Leiden: E. J. Brill, 1905.

Kaithi

Bihar High Court of Judicature. *Selection of Hindusthani Documents from the Courts of Bihar*, compiled by S. K. Das. Patna, Bihar: Superintendent, Government Printing, 1939.

Grierson, George A. *A Handbook to the Kaithi Character*. 2nd rev. ed. Calcutta: Thacker, Spink & Co., 1899. Revised edition of *A Kaithi Handbook*, 1881.

King, Christopher R. *One Language, Two Scripts: The Hindi Movement in Nineteenth Century North India*. Bombay: Oxford University Press, 1994.

Kayah Li

Bennett, J. Fraser. *Kayah Li Script: A Brief Description*. Urbana-Champaign: University of Illinois, 1993.

Karenni Literature Department. *Ka¹ya³lhi¹-Ku³la³ Nghôchozha³: The Modern Western Kayah Li-English Lexicon*. [Chiang Mai]: Payap University, 1994. [without tones = Kayalhi-Kula Nghôchozha]

Solnit, David B. *Eastern Kayah Li: Grammar, Texts, Glossary*. Honolulu: University of Hawai'i Press, 1997. ISBN 0-8248-1743-5.

Kharoshthi

Glass, Andrew. *A Preliminary Study of Kharosthi Manuscript Paleography*. 2000. Thesis (M.A.), University of Washington, 2000.

Glass, Andrew. "KharoDEhG Manuscripts: A Window on GandhFran Buddhism." *Nagoya Studies in Indian Culture and Buddhism*, 24 (2004): 129–152. ISSN 0285-7154.

Salomon, Richard. *Ancient Buddhist Scrolls from GandhZra: The British Library Kharosthi Fragments*. Seattle: University of Washington Press; London: British Library, 1999. ISBN 029597768X; 0295977698 (pbk.).

Latin

ISO/IEC 646:1991, *Information technology - ISO 7-bit coded character set for information interchange*.

ISO 5426:1983, *Extension of the Latin alphabet coded character set for bibliographic information interchange*.

ISO 6438:1983, *Documentation - African coded character set for bibliographic information interchange*.

ISO 6937:1994, *Information technology - Coded graphic character set for text communication - Latin alphabet*.

ISO/IEC 8859, *Information technology - 8-bit single-byte coded graphic character sets*

Part 1: Latin alphabet No. 1 (1998).

Part 2: Latin alphabet No. 2 (1999).

Part 3: Latin alphabet No. 3 (1999).

Part 4: Latin alphabet No. 4 (1998).

Part 9: Latin alphabet No. 5 (1999)

Part 10: Latin alphabet No. 6 (1998).

ISO/IEC 10367:1991, *Information technology - Standardized coded graphic character sets for use in 8-bit codes*.

ANSI X3.4-1986 American National Standards Institute. *Coded character set - 7-bit American national standard code*.

ANSI Z39.47-1985 American National Standards Institute. *Extended Latin alphabet coded character set for bibliographic use*.

LVS 18-92 Latvian National Centre for Standardization and Metrology *Libiesu kodu tabula ar 191 simbolu*.

Kuruch, Rimma Dmitrievna. *Saamsko-russkiy slovar'*. Moskva: Russkiy iazyk. 1985

Lepcha

Mainwaring, G. B. *A Grammar of the Rong (Lepcha) Language as it Exists in the Dorjeling and Sikim Hills*. Delhi: Daya Publishing House, 1985 (1876).

Plaisier, H. "A Brief Introduction to Lepcha Orthography and Literature." *Bulletin of Tibetology* 41:1 (2005), 7–24.

Plaisier H. *A Grammar of Lepcha*. Leiden: Brill, 2007. (Brill's Tibetan Studies Library, Languages of the Greater Himalayan Region 5).

Limbu

Bairagi Kaila, ed. *Limbu-Nepali-Angreji śabdakoś*. [Limbu-Nepali-English Dictionary.] Kathmandu: Royal Nepal Academy, [in press.]

Cemjonga, Imana Simha. *Yakthun-Pene-Mikphula Pancheka*. = *Limbu-Nepali-Angareji śabdakoś*. = *Limbu-Nepali-English Dictionary*. [Lekhaka] Imanasimha Cemajon. [Kathamandu]: Nepala Ekedemi [2018 vi., i.e., 1962]

Driem, George van. *A Grammar of Limbu*. Berlin, New York: Mouton de Gruyter, 1987. (Mouton grammar library, 4.) ISBN 0-89925-345-8. Appendix: *Anthology of Kiranti scripts*, pp. 550–558.

Shafer, Robert. *Introduction to Sino-Tibetan*. Wiesbaden: Harrassowitz, 1966–1974.

Sprigg, R. K. "Limbu Books in the Kiranti Script." In *International Congress of Orientalists (24th: 1957: Munich)*. *Akten des Vierundzwanzigsten Internationalen Orientalisten-Kongresses München 28. August bis 4. September 1957*, hrsg. von Herbert Franke. Wiesbaden: Deutsche Morgenländische Gesellschaft, in Kommission bei Franz Steiner Verlag, 1959.

ISO/IEC 10646:2014 (E)

Sprigg, R. K. [Review of van Driem (1987)]. *Bulletin of the School of Oriental and African Studies, University of London*, 52 (1989):1.163–165.

Subba, B. B. *Limbu, Nepali, English Dictionary*. Gangtok: Text Book Unit, Directorate of Education, Govt. of Sikkim, 1979 [i.e. 1980]. Cover title: Yakthun-Pene-Mikphula-panchekva.

Subba, B. B. *Yakthun hu?siŋlam* (“Limbu self-teaching method”) = Limbu akṣar gāiḍ (“Limbu letter guide”). Gangtok: Kwaliti Stores, 1991?

Yorñhāñ, Khel Rāj. *Limbū Nepālī śabdakoś*. [Lalitpur]: 2052 B.S. [i.e. 1995].

Linear B and Cypriot

Bennett, Emmett L. “Aegean Scripts.” In *The World’s Writing Systems*, edited by Peter T. Daniels and William Bright. New York: Oxford University Press, 1996. ISBN 0-19-507993-0.

Chadwick, John. *The Decipherment of Linear B*. 2nd ed. London: Cambridge University Press., 1967 [i.e. 1968].

Chadwick, John. *Linear B and Related Scripts*. Berkeley: University of California Press; [London]: British Museum, 1987. (*Reading the Past*, v. 1.) ISBN 0-520-06019-9.

Hooker, J. T. *Linear B: An Introduction*. Bristol: Bristol Classical Press, 1980. ISBN 0-906515-69-6. Corrected printing published 1983. ISBN 0-906515-69-6; 0-906515-62-9 (pbk.).

International Colloquium on Mycenaean Studies (3rd: 1961: Racine, WI). *Mycenaean Studies: Proceedings of the Third International Colloquium for Mycenaean Studies held at “Wingspread,” 4–8 September 1961*, edited by Emmett L. Bennett, Jr. Madison: University of Wisconsin Press, 1964.

Masson, Olivier. *Les Inscriptions chypriotes syllabiques: recueil critique et commenté*. Réimpr. augm. Paris: E. de Boccard, 1983.

Sampson, Geoffrey. *Writing Systems: A Linguistic Introduction*. Stanford, CA: Stanford University Press, 1985. ISBN 0-8047-1254-9. Also published: London, Hutchinson. ISBN 0-09-156980-X; 0-09-173051-1 (pbk.).

Ventris, Michael. *Documents in Mycenaean Greek*. 1st ed. by Michael Ventris and John Chadwick with a foreword by Alan J. B. Wace. 2nd ed. by John Chadwick. Cambridge: Cambridge University Press, 1973. ISBN 0-521-08558-6.

Lisu

Bya, Yuliya. *Li-su Tho Uh Ba Pa Pha Tso So Du (Lisu Alphabet Primer)*. Chiang Mai: Christian Literature Fellowship, 2000.

Xu, Lin, Yuzhang Mu, and Xingzhi Gai, eds. *Lisuyu Jianzhi (A Sketch of the Lisu Language)*. Beijing: The Ethnic Publishing House, 1986. (*Chinese Minority Language Sketch Series*.)

Yunnan Minority Language Commission, and Weixi Culture and Education Bureau, eds. *Li-su Tho Uh Tso So Du (Lisu Primer)*. Kunming: Yunnan Nationality Publishing House, 1981.

Zhu, Faqing. *Li-su Be Xuh Ngo Bae Khuh Tae Du Ra (Small Lisu-Chinese Dictionary)*. Dehong: Dehong Nationality Publishing House, 1984.

Lycian

Carruba, O. “La scrittura licia.” *Annali della scuola normale superiore di Pisa, classe di letter e filosofia*. 3rd series. 8 (1978):849–867.

Melchert, H. Craig. "Lycian." In *The Cambridge Encyclopedia of the World's Ancient Languages*, edited by Roger Woodard, 591–600. Cambridge: Cambridge University Press, 2004. ISBN-13: 978-0521562560.

Lydian

Gérard, Raphaël. *Phonétique et morphologie de la langue lydienne*. Louvain-la-Neuve: Peeters, 2005.

Gusmani, Roberto. *Lydisches Wörterbuch mit grammatischer Skizze und Inschriftensammlung*. Heidelberg: Carl Winter, 1964.

Melchert, H. Craig. "Lydian." In *The Cambridge Encyclopedia of the World's Ancient Languages*, edited by Roger Woodard, 601–608. Cambridge: Cambridge University Press, 2004. ISBN-13: 978-0521562560.

Mandaic

Daniels, "Aramaic Scripts for Aramaic Languages," in Daniels & Bright, eds., *The World's Writing Systems*, Oxford University Press, 1996, pp. 511-513 "Mandaic."

Häberl, "Iranian Scripts for Aramaic Languages: *The Origin of the Mandaic Script*," Bulletin of the American Schools of Oriental Research, No. 341 (Feb., 2006), pp. 53-62.

Coulmas, *The Blackwell Encyclopedia of Writing Systems*, Blackwell 1999, p. 320 "Mandean script."

Mathematical Symbols

ISO 6862, *Information and documentation - Mathematical coded character set for bibliographic information interchange*.

ANSI Y10.20-1988 American National Standards Institute. *Mathematic signs and symbols for use in physical sciences and technology*.

Mathematical Markup Language (MathML) Version 2.0. (W3C Recommendation 21 February 2001). Editors: David Carlisle, Patrick Ion, Robert Miner, [and] Nico Poppelier. Latest version: <http://www.w3.org/TR/MathML2/>

Selby, Samuel M. *Standard mathematical tables*. – 16th ed. – Cleveland, OH : Chemical Rubber Co., 1968. Shepherd, Walter.

STIPub Consortium. STIX (Scientific and Technical Information Exchange) Project. <http://www.ams.org/STIX/>

Swanson, Ellen. *Mathematics into Type*. Updated ed. by Arlene O'Sean and Antoinette Schleyer. Providence, RI: American Mathematical Society, 1999. ISBN 0-8218-1961-5.

Meetei Mayek

Chelliah, Shobhana L. *A Grammar of Meithei*. Berlin and New York: Mouton de Gruyter, 1997. ISBN 978-3-11-014321-8.

Debendra Singh, N. *Evolution of Manipuri Script*. [Imphal]: Manipur University, Centre for Manipuri Studies, 1990. (Research Report, 5).

Meroitic

Griffith, F. LI. Karanòg: *The Meroitic inscriptions of Shablûl and Karanòg*. Philadelphia: University Museum, 1911.

Millet, N. B. "The Meroitic script." In *The World's Writing Systems*, edited by Peter T. Daniels and William Bright. New York: Oxford University Press, 1996. ISBN 0-19-507993-0.

ISO/IEC 10646:2014 (E)

Rilly, Claude. *La langue du royaume de Méroé: un panorama de la plus ancienne culture écrite d'Afrique subsaharienne*. Paris: Librairie Honoré Champion, 2007.

Miao

Enwall, Joakim. *A Myth Become Reality: History and development of the Miao written language*. 2 vols. Stockholm: Institute of Oriental Languages, Stockholm University, 1994–1995. (Stockholm East Asian monographs no. 5-6.)

Xiong Yuyou. *Miao zu wen hua shi = A Cultural History of the Miao Nationality*. Kunming Shi: Yunnan min zu chu ban she, 2003.

Musical Symbols

ELOT 1373. *The Greek Byzantine Musical Notation System*. Athens, 1997 (ΣΕΠ ΕΛΟΤ 1373: 1997).

Catholic Church. *Graduale Sacrosanctae Romanae Ecclesiae de Tempore et de Sanctis SS. D. N. Pii X. Pontificis Maximi*. Parisiis: Desclée, 1961. (Graduale Romanum, no. 696.)

Gazimihal, Mahmut R. *Anadolu türküleri ve mûsikî istikbâlimiz* [by] Mahmut Ragip. [Istanbul]: Mârifet Matbaası, 1928.

Heussenstamm, George. *Norton Manual of Music Notation*. New York: W.W. Norton, 1987. ISBN 0-393-95526-5 (pbk.).

Kennedy, Michael. *Oxford Dictionary of Music*. Oxford, New York: Oxford University Press, 1985. ISBN 0-19-311333-3. Second ed. published 1994. ISBN 0-19-869162-9.

New Encyclopedia Britannica. 15th ed. Entry for “Music.”

The New Harvard Dictionary of Music, edited by Don Michael Randel. Cambridge, MA: Belknap Press of Harvard University Press, 1986. ISBN 0-674-61525-5.

Ottman, Robert W. *Elementary Harmony: Theory and Practice*. 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1970. ISBN 0-13-257451-9. Fifth ed. published 1998. ISBN 0-13-281610-5.

Rastall, Richard. *The Notation of Western Music: An Introduction*. London: Dent, 1983. ISBN 0-460-04205-X. Also published: New York: St. Martin's Press, 1982. ISBN 0-312-57963-2.

Read, Gardner. *Music Notation: A Manual of Modern Practice*. Boston: Allyn and Bacon, 1964.

Stone, Kurt. *Music Notation in the Twentieth Century: A Practical Guidebook*. New York: W.W. Norton, 1980. ISBN 0-393-95053-0.

Understanding Music with AI: Perspectives on Music Cognition, edited by Mira Balaban, Kemal Ebcioglu, and Otto Laske. Cambridge, MA: MIT Press; Menlo Park, CA: AAAI Press, 1992. ISBN 0-262-52170-9.

Myanmar

Mranmā–Aṅglip abhidhān = Myanmar–English Dictionary. Rankun: Dept. of Myanmar Language Commission, Ministry of Education, Union of Myanmar, 1993. Compiled and edited by the Myanmar Language Commission.

Mranmā cālui:poṅg:satpui kyam: nhañ. khwaithā.. [Rankun]: 1996. Translated title: Myanmar orthography treatise.

Okell, John. 1971. *A guide to the romanization of Burmese*. (James G. Forlang Fund; 27) London: Royal Asiatic Society of Great Britain and Ireland.

Roop, D. Haigh. *An Introduction to the Burmese Writing System*. [Honolulu]: Center for Southeast Asian Studies, University of Hawaii at Manoa, 1997. (Southeast Asia Paper, 11). Originally published: New Haven: Yale University Press, 1972. (Yale linguistic series). ISBN 0-300-01528-3.

N'Ko

Kanté, Souleymane. *Méthode pratique d'écriture n'ko*, 1961. Kankan, Guinea: Association de tradithérapeutes et pharmacologues, 1995.

N'Ko: The Common Language of Mandens. www.nkoinstitute.com

N'Ko: The Mandingo Language Site. www.kanjamadi.com

Ogham

I. S. 434:1999, *Information Technology - 8-bit single-byte graphic coded character set for Ogham = Teicneolaíocht Eolais - Tacar carachtar grafach Oghaim códaithe go haonbheartach le 8 ngiotán*. National Standards Authority of Ireland.

McManus, Damian. *A Guide to Ogam*. Maynooth: An Sagart, 1991. (Maynooth monographs, 4). ISBN 1-87068-417-6.

Oi Chiki

Hembram, S. M., et al. *Adibasi Oi script = at'ip'asi al ciki*. Calcutta: Adibasi Socio-Educational & Cultural Association, 1972.

Murmu, Raghunath. Ranar: *A Santali Grammar in Santali*. Singhbhum, Bihar: Adibasi Socio-Educational & Cultural Association, 1972.

Zide, Norman. "Scripts for Munda languages." In *The World's Writing Systems*, edited by Peter T. Daniels and William Bright. New York; Oxford: Oxford University Press, 1996. ISBN 0-19-507993-0.

Old Italic

Bonfante, Larissa. "The Scripts of Italy." In *The World's Writing Systems*, edited by Peter T. Daniels and William Bright. New York: Oxford University Press, 1996. ISBN 0-19-507993-0.

Cristofani, Mauro. "L'alfabeto etrusco." In *Lingue e dialetti dell'Italia antica*, a cura di Aldo Larosdocimi. Roma: Biblioteca di storia patria, a cura dell' Ente per la diffusione e l'educazione storia, 1978. (*Popoli e civiltà dell'Italia antica*, VI.)

Gordon, Arthur E. *Illustrated Introduction to Latin Epigraphy*. Berkeley: University of California Press, 1983. ISBN 0-520-03898-3.

Marinetti, Anna. *Le iscrizioni sudpicene*. I. Testi. Firenze: Olschki, 1985. ISBN 88-222-3331-X (v. 1).

Parlangeli, Oronzo. *Studi Messapici*. Milano: Istituto lombardo di scienze e lettere, 1960.

Old Persian

Schmitt, Rüdiger. *The Bisitun Inscriptions of Darius the Great, Old Persian Text*. London, School of Oriental and African Studies, 1991 (*Corpus Inscriptionum Iranicarum*, Part I: Inscriptions of ancient Iran, v.1, Text 1). ISBN 0-7286-0181-8.

Schweiger, Günter. *Kritische Neuedition der achaemenidischen Keilinschriften*. Taimering: Schweiger VWT-Verlag, 1998. (*Studien zur Iranistik*). ISBN 3-934548-00-8.

Old South Arabian

ISO/IEC 10646:2014 (E)

Nebes, Norbert, and Peter Stein. "Ancient South Arabian." In *The Cambridge Encyclopedia of the World's Ancient Languages*, edited by Roger D. Woodard. 454–487. Cambridge University Press, 2004. ISBN-13: 978-0521562560.

Ryckmans, J. "Origin and Evolution of South Arabian Minuscule Writing on Wood (1)." *Arabian Archaeology and Epigraphy* 12 (2001): 223–235. ISSN 0905-7196.

Smithsonian Institution. "Written in Stone: Inscriptions from the National Museum of Saudi Arabia." http://www.mnh.si.edu/epigraphy/figs-stones/x-large/color_xl_jpeg/fig02.jpg

Stein, Peter. "The Ancient South Arabian Minuscule Inscriptions on Wood: A New Genre of Pre-Islamic Epigraphy." *Jaarbericht van het Vooraziatisch-Egyptisch Genootschap "Ex Oriente Lux"*, 39 (2005): 181–199. ISSN 0075-2118.

Old Turkic

Erdal, Marcel. *A Grammar of Old Turkic*. Leiden & Boston: Brill, 2004. ISBN 9004102949.

Scharlipp, Wolfgang Ekkehard. *Eski Türk run yazıtlarına giri!: ders kitabı = An Introduction to the Old Turkish Runic Inscriptions: A Textbook in English and Turkish*. Engelschoff: Auf dem Ruffel, 2000. ISBN 3-933847-00-X.

von Gabain, A. *Alttürkische Grammatik mit Bibliographie, Lesestücken und Wörterverzeichnis, auch Neutürkisch*. Leipzig: Harrassowitz, 1941. (Porta Linguarum Orientalium, 23).

Osmanya

Afkeenna iyo fartiisa: buug koowaad. Xamar: Goosanka afka iyo suugaanta Soomaalida, 1971. Translated title: *Our language and its handwriting: book one*.

Cerulli, Enrico. "Tentativo indigeno di formare un alfabeto somalo." *Oriente moderno*, 12 (1932): 212–213. ISSN 0030-5472.

Gaur, Albertine. *A History of Writing*. London: British Library, 1992. ISBN 0-7123-0270-0. Also published: Rev. ed. New York: Cross River Press, 1992. ISBN 1-558-59358-6.

Gregersen, Edgar A. *Language in Africa: An Introductory Survey*. New York: Gordon and Breach, 1977. (Library of Anthropology). ISBN: 0-677-04380-5; 0-677-04385-6 (pbk.).

Maino, Mario. "L'alfabeto 'Osmania' in Somalia." *Rassegna di studi etiopici*, 10 (1951): 108–121. ISSN 0390-3699.

Nakanishi, Akira. *Writing Systems of the World: Alphabets, Syllabaries, Pictograms*. Rutland, VT: Tuttle, 1980. ISBN 0-8048-1293-4; 0-8048-1654-9 (pbk.). Revised translation of Sekai no moji.

Phags-pa

Luo, Changpei. *Basibazi yu Yuandai Hanyu [ziliao huibian] / Luo Changpei, Cai Meibiao bian zhu*. Beijing: Kexue chubanshe, 1959.

Poppe, Nikolai Nikolaevich. *The Mongolian Monuments in hP'ags-pa Script*. Translated and edited by John R. Krueger. 2nd ed. Wiesbaden: Harrassowitz, 1957. (Göttinger asiatische Forschungen, 8).

Zhaonasiu. *Menggu ziyun jiaoben / Zhaonasiu, Yang Naisi bian zhu*. [Beijing]: Min zu chu ban she, 1987. Author Zhaonasiu also known as Jagunasutu or Junast.

Philippines Scripts

Doctrina Christiana: *The First Book Printed in the Philippines, Manila 1593*. A facsimile of the copy in the Lessing J. Rosenwald Collection, with an introductory essay by Edwin Wolf II. Washington, DC: Library of Congress, 1947.

Kuipers, Joel C., and Ray McDermott. "Insular Southeast Asian Scripts." In *The World's Writing Systems*. Edited by Peter T. Daniels and William Bright. New York: Oxford University Press, 1996. ISBN 0-19-507993-0.

Santos, Hector. *The Living Scripts*. Los Angeles: Sushi Dog Graphics, 1995. (Ancient Philippine scripts series, 2). User's guide accompanying Computer Fonts, Living Scripts software.

Santos, Hector. *Our Living Scripts*. January 31, 1997. <http://www.bibingka.com/dahon/living/living.htm> Part of his A Philippine Leaf.

Santos, Hector. *The Tagalog Script*. Los Angeles: Sushi Dog Graphics, 1994. (Ancient Philippine scripts series, 1). User's guide accompanying Tagalog Script Fonts software.

Santos, Hector. *The Tagalog Script*. October 26, 1996.
<http://www.bibingka.com/dahon/tagalog/tagalog.htm> Part of his A Philippine Leaf.

Phoenician

Branden, Albertus van den. *Grammaire phénicienne*. Beyrouth: Librairie du Liban, 1969. (Bibliothèque de l'Université Saint-Esprit, 2).

McCarter, P. Kyle. *The Antiquity of the Greek Alphabet and the Early Phoenician Scripts*. Missoula, MT: Published by Scholars Press for Harvard Semitic Museum, 1975. (Harvard Semitic Monographs; 9.) ISBN 0-89130-066-X.

Noldeke, Theodor. *Beiträge zur semitischen Sprachwissenschaft*. Strassburg: Karl J. Trübner, 1904. Reprinted as: vol. 1 of *Beiträge und Neue Beiträge zur semitischen Sprachwissenschaft: achtzehn Aufsätze und Studien*. Amsterdam: APA-Philo Press, [1982]. Also published on microfiche by the American Theological Library Association.

Powell, Barry B. *Homer and the Origin of the Greek Alphabet*. Cambridge, New York: Cambridge University Press, 1991. ISBN 0-521-37157-0. Reprinted, 1996. ISBN 0-521-58907-X (pbk).

Rejang

Jaspan, M. A. *Folk Literature of South Sumatra: Redjang Ka-Ga-Nga Texts*. Canberra: Australian National University, 1964.

Runic

Benneth, Solbritt, Jonas Ferenius, Helmer Gustavson, & Marit Åhlén. 1994. *Runmärkt: från brev till klotter. Runorna under medeltiden*. [Stockholm]: Carlsson Bokförlag. ISBN 91-7798-877-9

Derolez, René. 1954. *Runica manuscripta: the English tradition*. (Rijksuniversiteit te Gent: Werken uitgegeven door de Faculteit van de Wijsbegeerte en Letteren; 118e aflevering) Brugge: De Tempel.

Friesen, Otto von. *Runorna*. Stockholm, A. Bonnier [1933]. (Nordisk kultur, 6).

Haugen, Einar Ingvald. *The Scandinavian Languages: An Introduction to Their History*. London: Faber, 1976. ISBN 0-571-10423-1. Also published: Cambridge, MA: Harvard University Press, 1976. ISBN 0-674-79002-2.

Musset, Lucien. *Introduction à la runologie*. Paris: Aubier-Montaigne, 1965.

ISO/IEC 10646:2014 (E)

Page, Raymond Ian. *Runes*. Berkeley: University of California Press; [London]: British Museum, 1987. (Reading the Past). ISBN 0-520-06114-4. British Museum Publications edition has ISBN 0-7141-8065-3.

Samaritan

Ben-Hayyam, Ze'ev. *A Grammar of Samaritan Hebrew, Based on the Recitation of the Law in Comparison with the Tiberian and other Jewish Traditions*. Jerusalem: Hebrew University Magnes Press, 2000. ISBN 1-57506-047-7.

Macuch, Rudolf. *Grammatik des samaritanischen Hebräisch*. Berlin: Walter de Gruyter, 1969. ISBN 9783110083767.

Murtonen, A. *Materials for a Non-Masoretic Hebrew Grammar III: A Grammar of the Samaritan Dialect of Hebrew*. Helsinki: Societas Orientalis Fennica, 1964. (Studia Orientalia, 29).

Saurashtra

Norihiko Učida. *Language of the Saurashtrians in Tirupati*. 2nd revised ed. Bangalore: Mahalaxmi Enterprises, 1991. (In Latin script.)

Norihiko Učida. *Saurashtra-English Dictionary*. Wiesbaden: Harrassowitz, 1990. ISBN 3 447030550. (In Latin script.)

Sharada

Deambi, Kaul and Bushan Kumar. *Śāradā and Ṭākārī Alphabets: Origin and Development*. New Delhi: Indira Gandhi National Centre for the Arts, 2008.

Grierson, George A. "On the Sharada Alphabet." *The Journal of the Asiatic Society of Great Britain and Ireland*, (1916): 677–708.

Shavian

ConScript Unicode Registry [by] John Cowan and Michael Everson. "E700–E72F Shavian." Included in the ConScript Registry (<http://www.evertype.com/standards/csur/index.html>) in 1997. Shavian was withdrawn from the ConScript Registry in 2001, because of its addition to the Unicode Standard and ISO/IEC 10646.

Crystal, David. *The Cambridge Encyclopedia of Language*. Cambridge, New York: Cambridge University Press, 1987. ISBN 0-521-26438-3. 2nd ed. Cambridge, New York: Cambridge University Press, 1997. ISBN 0-521-55050-5; 0-521-55967-7.

DeMeyere, Ross. *About Shavian*. 1997. <http://www.demeyere.com/Shavian/info.html>.

Shaw, George Bernard. *Androcles and the Lion: An Old Fable Renovated, by Bernard Shaw, with a Parallel Text in Shaw's Alphabet to Be Read in Conjunction Showing Its Economies in Writing and Reading*. Harmondsworth: Penguin Books, 1962.

Sinhala

SLS 1134:1996 Sri Lanka Standards Institution *Sinhala character code for information interchange*.

Gunasekara, Abraham Mendis. *A comprehensive grammar of the Sinhalese language*. New Delhi: Asian Educational Services, 1986(Reprint of 1891 edition).

Sora Sompeng

Mahapatra, Khageshwar. "'Soraṅ Sompen': A Sora Script." Unpublished conference paper. Delhi, Mysore, 1978–1979.

Zide, Norman. "Scripts for Munda languages." In *The World's Writing Systems*, edited by Peter T. Daniels and William Bright. New York: Oxford University Press, 1996. ISBN 0-19-507993-0.

Zide, Norman. "Three Munda scripts." In *Linguistics of the Tibeto-Burman Area*. Vol. 22.2—Fall 1999

Sundanese

Baidillah, Idin, Cucu Komara, and Deuis Fitni. Ngalagena: *Panglengkep Pangajaran Aksara Sunda pikeun Murid Sakola Dasar/Dikdas 9 Taun*. [Bandung]: CV Walatra, [2002].

Hardjasaputra, A. Sobana, Tedi Permadi, Undang A. Darsa, and Edi S. Ekadjati. *Rancangan Pembakuan Aksara Sunda*. Bandung, 1998.

Symbols (Miscellaneous)

ISO 2033:1983, *Information processing - Coding of machine readable characters (MICR and OCR)*.

ISO 2047:1975, *Information processing - Graphical representations for the control characters of the 7-bit coded character set*.

ISO/IEC 9995-7:1994, *Information technology – Keyboard layouts for text and office systems – Part 7: Symbols used to represent functions*.

ANSI X3.32-1973 American National Standards Institute. *American national standard graphic representation of the control characters of American national standard code for information interchange*.

ANSI Y14.5M-1982 American National Standard. *Engineering drawings and related document practices, dimensioning and tolerances*.

Syriac

Kefarnissy, Paul. *Grammaire de la langue araméenne syriaque*. Beyrouth, 1962.

Nöldeke, Theodor. *Compendious Syriac Grammar*. With a table of characters by Julius Euting. Translated from the 2nd and improved German ed., by James A. Crichton. London: Williams & Norgate, 1904. Reprinted: Tel Aviv: Zion Pub. Co. [1970].

Robinson, Theodore Henry. *Paradigms and Exercises in Syriac Grammar*. 4th ed. Rev. by L. H. Brockington. Oxford: Clarendon Press; New York: Oxford University Press, 1962. ISBN 0-19-815416-X, 0-19-815458-5 (pbk.).

Tai Le

Coulmas, Florian. *The Blackwell Encyclopedia of Writing Systems*. Oxford, Cambridge: Blackwell, 1996. ISBN 0-631-19446-0. Dehong writing, pp. 118–119.

Lá a² maw³ lá a² ka va³ m² tse² law ya pa me na⁴ ka na: tá va ʔá na kó ma⁶ sá na² teh ma⁶. Yina⁵lána⁵ mina⁵su⁴ su⁴pána²se³ (Yunnan minzu chubanshe). 1988. ISBN 7-5367-1100-4.

Tsa va⁴ má³ hó va³: la ta⁶ mé² sá a³ seh va² xo ɲa³. Yina⁵lána⁵ mina⁵su⁴ su⁴pána²se³ (Yunnan minzu chubanshe). 1997. ISBN 7-5367-1455-6.

Tai Tham

Peltier, Anatole-Roger. 1996. *Lanna Reader*. Chiang Mai: Wat Tha Kradas.

Kasēm Siriratphiriya, and Mahāwitthayālai Sukhōthaihammāthirāt. *Tūa Mueang: kārīan phāsā Lānnā phān khōngsāng kham*. Nonthaburī: Rōngphim Mahāwitthayālai Sukhōthaihammāthirāt, 2548 [2005]. ISBN 974-9942-00-0.

ISO/IEC 10646:2014 (E)

Rungrueangsri, Udom. 2004. *Pacanānukrom Lānnā-Thai: Chabaph maefāhluang*. ISBN 974-685-175-9.
Baephryar phāsā Lānnā. ISBN 974-386-044-4.

Takri

Deambi, Kaul and Bushan Kumar. *Śāradā and Tākārī Alphabets: Origin and Development*. New Delhi: Indira Gandhi National Centre for the Arts, 2008.

Thaana

Geiger, Wilhelm. *Maldivian Linguistic Studies*. New Delhi: Asian Educational Services, 1996. ISBN 81-206-1201-9. Originally published: Colombo: H. C. Cottle, Govt. Printer, 1919.

Maniku, Hassan Ahmed. *Say It in Maldivian (Dhivehi)*, [by] H. A. Maniku [and] J. B. Disanayaka. Colombo: Lake House Investments, 1990.

Tibetan

Beyer, Stephen V. *The classical Tibetan language*. State University of New York. ISBN 0-7914-1099-4

Ugaritic

O'Connor, M. "Epigraphic Semitic Scripts." In *The World's Writing Systems*, edited by Peter T. Daniels and William Bright. New York: Oxford University Press, 1996. ISBN 0-19-507993-0.

Walker, C. B. F. *Cuneiform*. London: British Museum Press, 1987. (Reading the Past, v. 3.) ISBN 0-7141-8059-9. University of California Press edition has ISBN 0-520-06115-2 (pbk.).

Thai

TIS 620-2533 *Thai Industrial Standard for Thai Character Code for Computer*. (1990)

Vai

Dalby, David. "A Survey of the Indigenous Scripts of Liberia and Sierra Leone: Vai, Mende, Loma, Kpelle and Bassa." *African Language Studies* 8 (1967):1–51.

Kandakai, Zuke, et al. *Vai kpolo saikilamaa mɛ = The Standard Vai Script*. Monrovia: University of Liberia African Studies Program, 1962.

Massaquoi, Momolu. "The Vai People and Their Syllabic Writing." *Journal of the Royal African Society* 10.40, July (1911), 459–466.

Singler, John. "Scripts of West Africa." In *The World's Writing Systems*, edited by Peter T. Daniels and William Bright. New York: Oxford University Press, 1996. ISBN 0-19-507993-0.

Stewart, Gail, and P. E. H. Hair. "A Bibliography of the Vai Language and Script." *Journal of West African Languages* 6.2 (1969), 124.

Yi

GB13134: *Xinxi jiaohuanyong yiwen bianma zifuji (Yi coded character set for information interchange)*, [prepared by] Sichuansheng minzushiwu weiyuanhui. Beijing, Jishu Biaozhun Chubanshe (Technical Standards Press), 1991. (GB 13134-1991).

Nuo-su bbur-ma shep jie zzit. = Yi wen jian zi ben. Chengdu: Sichuan minzu chubanshe, 1984.

Nip huo bbur-ma ssix jie. = Yi Han zidian. Chengdu: Sichuan minzu chubanshe, 1990. ISBN 7-5409-0128-4.

Annex N (informative)

External references to character repertoires

N.1 Methods of reference to character repertoires and their coding

Within programming languages and other methods for defining the syntax of data objects there is commonly a need to declare a specific character repertoire from among those that are specified in this International Standard. There may also be a need to declare the corresponding coded representations applicable to that repertoire.

For any character repertoire that is in accordance with this International Standard a precise declaration of that repertoire should include the following parameters:

- identification of this International Standard,
- the adopted subset of the repertoire, identified by one or more collection numbers,
- the code unit sequence content definition,
- the adopted encoding form (UTF-8, UTF-16, or UTF-32).

One of the methods now in common use for defining the syntax of data objects is Abstract Syntax Notation 1 (ASN.1) specified in ISO/IEC 8824. The corresponding coded representations are specified in ISO/IEC 8825. When this method is used the forms of the references to character repertoires and coding are as indicated in the following clauses.

N.2 Identification of ASN.1 character abstract syntaxes

The set of all character strings that can be formed from the characters of an identified repertoire in accordance with this International Standard is defined to be a “character abstract syntax” in the terminology of ISO/IEC 8824. For each such character abstract syntax, a corresponding object identifier value is defined to permit references to be made to that syntax when the ASN.1 notation is used.

ISO/IEC 8824-1 Annex B specifies the form of object identifier values for objects that are specified in an ISO standard. In such an object identifier the features and options of this International Standard are identified by means of numbers (arcs) which follow the arcs “10646” and “0” which identify the whole ISO/IEC 10646.

NOTE 1 – The arc (0) is required to complement the arcs (1) and (2) which represent respectively ISO/IEC 10646-1 and ISO/IEC 10646-2. These two arcs should not be used.

The first such arc following a 10646 arc identifies the code unit sequence content definition, and is referred to as ‘level-3 (3)’.

NOTE 2 – This version of the standard specifies a single definition for code unit sequence content. That definition was formerly known as implementation level 3 in previous editions of this standard

The second such arc identifies the repertoire subset, and is either

- all (0), or
- collections (1).

Arc (0) identifies the entire collection of characters specified in this International Standard. No further arc follows this arc.

NOTE 3 – This collection includes private planes, and is therefore not fully-defined. Its use without additional prior agreement is deprecated.

Arc (1) is followed by one or a sequence of further arcs, each of which is a collection number from Annex A, in ascending numerical order. This sequence identifies the subset consisting of the collections whose numbers appear in the sequence.

ISO/IEC 10646:2014 (E)

NOTE 4 – As an example, the object identifier for the subset comprising the collections BASIC LATIN, LATIN-1 SUPPLEMENT, and MATHEMATICAL OPERATORS is:

{iso standard 10646 (0) level-3 (3) collections (1) 1 2 39}

ISO/IEC 8824 also specifies object descriptors corresponding to object identifier values. For an unrestricted repertoire, the corresponding object descriptor is as follows:

3 0 : "ISO 10646 level-3 unrestricted"

For a single collection with collection name "xxx".

3 1 : "ISO 10646 level-3 xxx"

For a repertoire comprising more than one collection, numbered m1, m2, etc.

3 1 : "ISO 10646 level-3 collections m1, m2, m3, .. "

NOTE 5 – All spaces are single spaces.

N.3 Identification of ASN.1 character transfer syntaxes

The coding method for character strings that can be formed from the characters in accordance with this International Standard is defined to be "character transfer syntax" in the terminology of ISO/IEC 8824. For each such character transfer syntax, a corresponding object identifier value is defined to permit references to be made to that syntax when the ASN.1 notation is used.

In an object identifier in accordance with ISO/IEC 8824-1 Annex B, the coded representation form specified in this International Standard is identified by means of numbers (arcs) which follow the arcs "10646" and "0" which identify the whole ISO/IEC 10646.

The first such arc is

transfer-syntaxes (0).

The second such arc identifies the encoding form and is either

four-octet-form (4) for the UTF-32 encoding form, or
utf16-form (5) for the UTF-16 encoding form, or
utf8-form (8) for the UTF-8 encoding form.

NOTE 1 – As an example, the object identifier for the UTF-32 encoding form is:

{iso standard 10646 (0) transfer-syntaxes (0) four-octet-form (4)}

The following object identifier is also valid but deprecated:

{iso standard 10646 (1) transfer-syntaxes (0) four-octet-form (4)}

NOTE 2 – Previous versions of this standard supported a two-octet-BMP-form (2) arc which is now deprecated.

The corresponding object descriptors are:

"ISO 10646 form 4"

"ISO 10646 utf-16"

"ISO 10646 utf-8".

NOTE 3 – Previous versions of this standard supported the "ISO 10646 form 2" object descriptor which is now deprecated.

Annex P (informative)

Additional information on CJK Unified Ideographs












Annex P contains additional information on CJK Unified Ideographs.

NOTE – The first edition of this International Standard (ISO/IEC 10646:2003 and amendments 1 to 5) used this annex to provide additional information on all characters. This edition of the standard includes most of that information in the code charts. Because the code charts for CJK unified ideographs do not include any name list, the information about these characters is still included here.

Each entry in the table P.1 consists for each row of an extract of the CJK Unified Ideograph code point entry in the code chart, followed in the next column by the related additional information. Entries are arranged in ascending sequence of code point.

Table P.1: Additional information on CJK Unified Ideographs

UCS / Glyph	Additional information
9FB9 𠂇 <small>八 12.4</small> G9-FE7E	These three characters are intended to represent a component at a specific position of a full ideograph. The ideographs representing the same structure without a preferred positional preference are encoded at 20509 𠂇, 2099D 𠂈, and 470C 𠂉 respectively.
9FBA 𠂈 <small>十 24.6</small> G9-FE90	
9FBB 𠂉 <small>言 149.12</small> G9-FEA0	
20885 𠂊 𠂋 𠂌 <small>力 19.10</small> UCS2003 GKX-0148.26 T5-3669	T5-3669 source glyph was mistakenly unified to this code point.
22936 𠂍 𠂎 𠂏 <small>心 61.16</small> UCS2003 GKX-0408.28 T5-6777	T5-6777 source glyph was mistakenly unified to this code point.
22BA3 𠂐 𠂑 𠂒 <small>手 64.8</small> UCS2003 GKX-0440.17 T6-492E	GKX-0440.17 source glyph was mistakenly unified to this code point.
23023 𠂓 𠂔 𠂕 <small>支 66.16</small> UCS2003 GKX-0476.21 T5-6C34	T5-6C34 source glyph was mistakenly unified to this code point.
235F1 𠂖 𠂗 <small>木 75.10</small> UCS2003 V0-354D	UCS2003 glyph for this code point was mistakenly designed.
2382C 𠂘 𠂙 <small>木 75.18</small> UCS2003 TF-6951	UCS2003 glyph for this code point was mistakenly designed.
23EE4 𠂚 𠂛 𠂜 <small>水 85.11</small> UCS2003 GKX-0648.09 T7-243F	T7-243F source glyph was mistakenly unified to this code point.
24229 𠂝 𠂞 𠂟 <small>火 86.7</small> UCS2003 GKX-0672.02 T4-3273	GKX-0672.02 source glyph was mistakenly unified to this code point.

UCS / Glyph	Additional information
24369 火 86.12  UCS2003 TF-5024	UCS2003 glyph for this code point was mistakenly designed.
243BE 火 86.12  UCS2003 T7-2F4B	The source glyph for T7-2F4B should have been unified with 24381 𤇀 but was allocated here by a mistake. The UCS2003 glyph for this code point should have been based on T7-2F4B but showed different shape by a mistake. For consistency with TCA CNS standards, 243BE's source reference to T7-2F4B is kept as in this International Standard.
24A8A 玉 96.13  UCS2003 V2-7C66	UCS2003 glyph for this code point was mistakenly designed.
24F15 疒 104.18  UCS2003 V2-7D5A	UCS2003 glyph for this code point was mistakenly designed.
25089 皿 108.10  UCS2003 V2-7D6B	UCS2003 glyph for this code point was mistakenly designed.
25B88 竹 118.8  UCS2003 V3-364B	UCS2003 glyph for this code point was mistakenly designed.
27555 虫 142.17  UCS2003 GKX-1103.29 T5-7649	UCS2003 glyph for this code point was mistakenly designed.
27B1F 言 149.12  UCS2003 GHZ-64018.09 T7-5035	GHZ-64018.09 source glyph was mistakenly unified to this code point.
27D41 貝 154.4  UCS2003 TF-385F	UCS2003 glyph for this code point was mistakenly designed.
28321 車 159.8  UCS2003 GKX-1244.18 T6-632A	T6-632A source glyph was mistakenly unified to this code point.
28599 辵 162.11  UCS2003 T5-516D V4-5565	V4-5565 source glyph was mistakenly unified to this code point.
28B75 金 167.13  UCS2003 TF-686D	The glyph of TF-686D in TCA CNS standard has been changed after the original publication of CJK UNIFIED IDEOGRAPHES EXTENSION B in ISO/IEC 10646. For consistency with TCA CNS standard, TF-686D glyph needs to be as in this International Standard, although the glyph is not usually unified with UCS2003 glyph of this code point.
293FB 韋 178.20  UCS2003 GHZ-74512.13 T5-7C22	The glyph of T5-7C22 in TCA CNS standard has been changed after the original publication of CJK UNIFIED IDEOGRAPHES EXTENSION B in ISO/IEC 10646. For consistency with TCA CNS standard, T5-7C22 glyph needs to be as in this International Standard, although the glyph is not usually unified with GHZ-74512.13 glyph and/or UCS2003 glyph of this code point.

UCS / Glyph	Additional information
<p>299FB 马 187.8</p> <p>𠂔 UCS2003 𠂔 GCH</p>	<p>The GCH glyph for this code point has been changed after the original publication of CJK UNIFIED IDEO GRAPHS EXTENSION B in ISO/IEC 10646. The GCH glyph needs to be as in this International Standard, although the glyph is not usually unified with UCS2003 glyph of this code point.</p>
<p>29C52 𪛗 193.12</p> <p>𪛗 UCS2003 𪛗 T7-5666</p>	<p>The glyph of T7-5666 in TCA CNS standard has been changed after the original publication of CJK UNIFIED IDEO GRAPHS EXTENSION B in ISO/IEC 10646. For consistency with TCA CNS standard, T7-5666 glyph needs to be as in this International Standard, although the glyph is not usually unified with UCS2003 glyph of this code point</p>
<p>2A0B8 鳥 196.9</p> <p>𪛗 UCS2003 𪛗 GKX-1494.15 𪛗 T7-523A</p>	<p>The source glyph of T7-523A in TCA CNS standard has been changed after the original publication of CJK UNIFIED IDEOGRAPHS EXTENSION B in ISO/IEC 10646-2. For consistency with TCA CNS standard, T7-523A glyph needs to be as in this International Standard, although the glyph is not usually unified with GKX-1494.15 glyph and/or UCS2003 glyph of this code point.</p>
<p>2A6C0 龜 213.8</p> <p>𪛗 UCS2003 𪛗 GKX-1538.20 𪛗 T5-7B5E</p>	<p>GKX-1538.20 source glyph was mistakenly unified to this code point.</p>

Annex Q
(informative)
Code mapping table for Hangul syllables

NOTE – The information concerning mapping between Hangul syllables (and code points) that were specified in the first edition of ISO/IEC 10646-1 and their amended code points is available in previous editions of this standard.

Annex R
(informative)
Names of Hangul syllables

Annex R provides the full name and annotation of Hangul syllables through a linked file:

The content linked to is a plain text file, using ISO/IEC 646-IRV characters with CARRIAGE RETURN/LINE FEED as end of line mark that specifies, after a 5-lines header, as all the Hangul syllables, each line specified as follows:

- 01-04 octet: code point in hexadecimal notation,
- 05 octet: SPACE character,
- 06 octet until end of line: Hangul syllable with the annotation between parenthesis.

[Click on this highlighted text to access the file containing the Hangul syllable names.](#)

NOTE – The content is also available as a separate viewable file in the same directory as this document. The file is named: "HangulSy.txt".

Annex S (informative)

Procedure for the unification and arrangement of CJK Ideographs

The graphic character collections of CJK unified ideographs in this International Standard are specified in Clause 31. They are derived from many more ideographs which are found in various different national and regional standards for coded character sets (the "sources").

Annex S describes how the ideographs in this standard are derived from the sources by applying a set of unification procedures. It also describes how the ideographs in this standard are arranged in the sequence of consecutive code points to which they are assigned.

The source references for CJK unified ideographs are specified in Clause 23.

Within the context of this International Standard a unification process is applied to the ideographic characters taken from the codes in the source groups. In this process, single ideographs from two or more of the source groups are associated together, and a single code point is assigned to them in this standard. The associations are made according to a set of procedures that are described below. Ideographs that are thus associated are described here as "unified".

NOTE – The unification process does not apply to the following collections of ideographic characters:

CJK RADICALS SUPPLEMENT (2E80 - 2EFF)

KANGXI RADICALS (2F00 - 2FDF)

CJK COMPATIBILITY IDEOGRAPHS (F900 - FAFF with the exception of FA0E, FA0F, FA11, FA13, FA14, FA1F, FA21, FA23, FA24, FA27, FA28 and FA29)

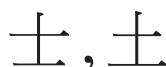
CJK COMPATIBILITY IDEOGRAPHS SUPPLEMENT (2F800-2FA1F).

S.1 Unification procedure

S.1.1 Scope of unification

Ideographs that are unrelated in historical derivation (non-cognate characters) have not been unified.

EXAMPLE



NOTE – The difference of shape between the two ideographs in the above example is in the length of the lower horizontal line. This is considered an actual difference of shape. Furthermore these ideographs have different meanings. The meaning of the first is "Soldier" and of the second is "Soil or Earth".

An association between ideographs from different sources is made here if their shapes are sufficiently similar, according to the following system of classification.

S.1.2 Two level classification

A two-level system of classification is used to differentiate (a) between abstract shapes and (b) between actual shapes determined by particular typefaces. Variant forms of an ideograph, which can not be unified, are identified based on the difference between their abstract shapes.

S.1.3 Procedure

A unification procedure is used to determine whether two ideographs have the same abstract shape or different ones. The unification procedure has two stages, applied in the following order:

- a) Analysis of component structure;
- b) Analysis of component features;

S.1.3.1 Analysis of component structure

In the first stage of the procedure the component structure of each ideograph is examined. A component of an ideograph is a geometrical combination of primitive elements. Alternative ideographs can be configured from the same set of components. Components can be combined to create a new component with a more complicated structure. An ideograph, therefore, can be defined as a component tree, where the top node is the ideograph itself, and the bottom nodes are the primitive elements. This is shown in Figure S.1.

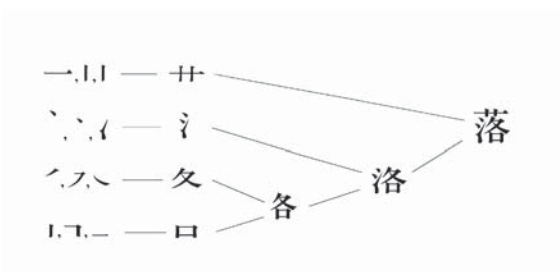


Figure S.1 - Component structure

S.1.3.2 Analysis of component features

In the second stage of the procedure, the components located at corresponding nodes of two ideographs are compared, starting from the top level, as shown in Figure S.2.

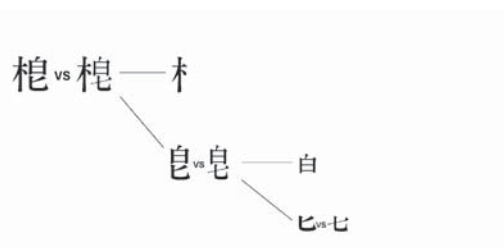


Figure S.2 - The most superior node of a component

The following features of each ideograph to be compared are examined:

- the number of components,
- the relative position of the components in each complete ideograph,
- the structure of corresponding components.

If one or more of the features a) to c) above are different between the ideographs in the comparison, the ideographs are considered to have different abstract shapes and are therefore not unified.

If all of the features a) to c) above are the same between the ideographs, the ideographs are considered to have the same abstract shape and are therefore unified.

S.1.4 Examples of differences of abstract shapes

To illustrate rules derived from a) to c) in S.1.3.2, some typical examples of ideographs that are not unified, owing to differences of abstract shapes, are shown below.

S.1.4.1 Different number of components

The examples below illustrate rule a) since the two ideographs in each pair have different numbers of components.

崖・厓, 肱・肱, 降・夆

S.1.4.2 Different relative positions of components

The examples below illustrate rule b). Although the two ideographs in each pair have the same number of components, the relative positions of the components are different.

峰·峯, 荊·荆

When the ideographs consists of two horizontally aligned components, a difference of the last stroke of the left-hand component going beneath the right-hand component should not warrant separate encoding, as in the case of the source glyphs for U+34F3:

匙·匙

S.1.4.3 Different structure of a corresponding component

The examples below illustrate rule c). The structure of one (or more) corresponding components within the two ideographs in each pair is different.

扌·扌, 策·箒, 𠂇·𠂇, 圣·𠂇, 𠂇·𠂇, 区·區, 夹·夾,
 单·單, 隹·隹, 𠂇·𠂇, 𠂇·𠂇, 𠂇·𠂇, 𠂇·𠂇, 𠂇·𠂇, 𠂇·𠂇,
 朵·朵, 隹·隹, 恒·恆, 𠂇·𠂇, 𠂇·𠂇, 𠂇·𠂇, 𠂇·𠂇

S.1.5 Differences of actual shapes

To illustrate the classification described in S.1.2, some typical examples of ideographs that are unified are shown below. The two or more ideographs in each group below have different actual shapes, but they are considered to have the same abstract shape, and are therefore unified on their own or as component in larger ideographs. The differences are classified according to the following examples.

a) Differences in rotated strokes/dots

半·半, 勺·勺, 羽·羽, 𠂇·𠂇, 兼·兼, 益·益, 每·每

b) Differences in overshoot at the stroke initiation and/or termination

身·身, 雪·雪, 拐·拐, 不·不, 非·非, 周·周, 告·告

c) Differences in contact of strokes

奧·奧, 酉·酉, 兕·兕, 查·查, 奔·奔

d) Differences in protrusion at the folded corner of strokes

巨·巨, 成·成

e) Differences in bent strokes

西·西

f) Differences in folding back at the stroke termination

朱·朱

g) Differences in accent at the stroke initiation

父·父, 丈·丈, 夂·夂

h) Differences in "rooftop" modification

八·八, 宀·宀

i) Addition or omission of a minor stroke

步·步, 者·者, 臭·臭, 呂·呂, 单·单

j) Combinations of the above differences

刃·刃, 直·直, 鼎·鼎

k) Miscellaneous

乂·乂·乂, 示·示·示, 艮·艮·艮, 食·食·食, 黃·黃, 盥·盥, 曷·曷, 包·包
 青·青, 册·册, 爭·爭, 畚·畚, 录·录, 并·并, 骨·骨, 吳·吳·吳,
 眞·眞·眞, 爲·爲, 曾·曾·曾, 專·專, 內·內, 晉·晉, 龜·龜, 𠂇·𠂇

NOTE – Some of the group items are unified when used as components in more complex ideographs, but are not unified themselves for other reasons, such as the source separation rule.

These differences in actual shapes of a unified ideograph are presented in the corresponding source columns for each code point entry in the code charts in Clause 31 of this International Standard.

S.1.6 Source separation rule

To preserve data integrity through multiple stages of code conversion (commonly known as “round-trip integrity”), any ideographs that are separately encoded in any one of the source standards listed below have not been unified.

G-source:	GB2312-80, GB12345-90, GB7589-87*, GB7590-87*, GB8565-88*, General Purpose Hanzi List for Modern Chinese Language*
T-source:	TCA-CNS 11643-1986/1st plane, TCA-CNS 11643-1986/2nd plane, TCA-CNS 11643-1986/14th plane*
J-source:	JIS X 0208-1990, JIS X 0212-1990
K-source:	KS X 1001:2004 (previously KS C 5601-1989), KS X 1002:2001 (previously KS C 5657-1991)

NOTE 1 – A “*” after the reference number of a standard indicates that some of the ideographs included in that standard are not introduced into the unified collection.

However, some ideographs encoded in two standards belonging to the same source group (e.g. GB2312-80 and GB12345-90) have been unified during the process of collecting ideographs from the source group.

The source separation rule described in S.1.6 only applies to the CJK UNIFIED IDEOGRAPHS block specified in the Basic Multilingual Plane.

NOTE 2 – CJK Compatibility Ideographs are created following a rule very similar to the source separation rule. However, the end result is the combination of a single CJK Unified Ideograph and one or several CJK Compatibility Ideographs. When the source separation rule is applied, all ‘similar’ source CJK Ideographs result in separate CJK Unified Ideographs.

S.2 Arrangement procedure

S.2.1 Scope of arrangement

The arrangement of the CJK UNIFIED IDEOGRAPHS in the code charts of Clause 31 of this International Standard is based on the filing order of ideographs in the following dictionaries.

Priority	Dictionary	Edition
1	Kangxi Dictionary 康熙字典	Beijing 7 th edition
2	Daikanwa Jiten 大漢和辭典	9 th edition
3	Hanyu Dazidian 漢語大字典	1 st edition
4	Daejaweon 大字源	1 st edition

The dictionaries are used according to the priority order given in the table above. Priority 1 is highest. If an ideograph is found in one dictionary, the dictionaries of lower priority are not examined.

S.2.2 Procedure

S.2.2.1 Ideographs found in the dictionaries

- If an ideograph is found in the Kangxi Dictionary, it is positioned in the code chart in accordance with the Kangxi Dictionary order.
- If an ideograph is not found in the Kangxi Dictionary but is found in the Daikanwa Jiten, it is given a position at the end of the radical-stroke group under which is indexed the nearest preceding Daikanwa Jiten character that also appears in the Kangxi dictionary.
- If an ideograph is found in neither the Kangxi nor the Daikanwa, the Hanyu Dazidian and the Daejaweon dictionaries are referred to with a similar procedure.

S.2.2.2 Ideographs not found in the dictionaries

If an ideograph is not found in any of the four dictionaries, it is given a position at the end of the radical-stroke group (after the characters that are present in the dictionaries) and it is indexed under the same radical-stroke count.

S.3 Source separation examples

The pairs (or triplets) of ideographs shown below are exceptions to the unification rules described in S.1. They are not unified because of the source separation rule described in S.1.6.

NOTE – The particular source group (or groups) that causes the source separation rule to apply is indicated by the letter (G, J, K, or T) that appears to the right of each pair (or triplet) of ideographs. The source groups that correspond to these letters are identified in S.1.6.

丟丟 4E1F 4E22	T	俁俁 4FC1 4FE3	TJK	兌兌 514C 5151	T	刃刃 5203 5204	TJ
么么 4E48 5E7A	GT	兪兪 4FDE 516A	T	兔兔 514E 5154	TJ	刊刊 520A 520B	TJ
争爭 4E89 722D	GTJ	俱俱 4FF1 5036	T	兗兗 5156 5157	T	刪刪 5220 522A	T
仞仞 4EDE 4EED	J	值值 5024 503C	T	冊冊 518A 518C	TJ	別別 5225 522B	T
併併 4F75 5002	T	偷偷 5077 5078	T	淨淨 51C0 51C8	G	券券 5238 52B5	TJ
侶侶 4FA3 4FB6	T	偽偽 507D 50DE	TJ	尢尢 51E2 51E3	T	剎剎 5239 524E	T

𪛗𪛗	T	告告	T	增增	T	媿媿	T
524F 5259		543F 544A		5897 589E		5A7E 5AAE	
剝剝	T	唧唧	T	壯壯	GTJ	媼媼	TK
525D 5265		5527 559E		58EE 58EF		5AAA 5ABC	
劒劒	J	喻喻	T	壽壽	T	媯媯	T
5292 5294		55A9 55BB		58FD 5900		5AAF 5B00	
勻勻	T	噓噓	T	𩇛𩇛	T	嫫嫫	T
52FB 5300		5618 5653		5910 657B		5B0E 5B14	
单单	T	噠噠	GTJ	夨夨	GTJ	嫫嫫	GT
5355 5358		568F 5694		5932 672C		5B24 5B37	
卽卽	TK	圀圀	T	奧奧	J	孳孳	T
5373 537D		56EF 56FD		5965 5967		5B73 5B76	
卷卷	TJ	圈圈	TJ	獎獎獎	TJ	宮宮	T
5377 5DFB		5708 570F		5968 596C 734E		5BAB 5BAE	
叁叁	GT	圓圓	T	妝妝	GT	寬寬	T
53C1 53C2		570E 5713		5986 599D		5BDB 5BEC	
參參	T	圖圖	T	妍妍	T	寧寧	T
53C3 53C4		5716 5717		598D 59F8		5BDC 5BE7	
呂呂	T	𠂔𠂔	T	姍姍	T	寢寢	GTJ
5415 5442		5759 5DE0		59CD 59D7		5BDD 5BE2	
吞吞	T	埤埤	J	姪姪	GT	專專	J
541E 5451		57D2 57D3		59EB 59EC		5C02 5C08	
吳吳吳	TJ	塹塹	T	娛娛娛	T	將將	GTJ
5433 5434 5449		5848 588D		5A1B 5A2F 5A31		5C06 5C07	
訥訥	T	塡塡	TJ	婕婕	T	尔尔	T
5436 5450		5861 586B		5A55 5AAB		5C13 5C14	

尙尙	T	𠂇𠂇	TJ	慎慎	TJ	搵搵	T
5C19 5C1A		5F50 5F51		613C 614E		63FE 6435	
尙尙	T	𠂇𠂇	T	戩戩	GT	擊擊	TJ
5C2A 5C2B		5F54 5F55		6229 622C		6483 64CA	
𨮒𨮒	T	彙彙	T	戲戲	T	教教	T
5C36 5C37		5F59 5F5A		622F 6231		654E 6559	
屏屏	T	彝彝	J	戶戶戶	T	斂斂	T
5C4F 5C5B		5F5B 5F5C		6236 6237 6238		6553 655A	
崢崢	GT	彝彝	T	戾戾	T	既既	T
5CE5 5D22		5F5D 5F5E		623B 623E		65E2 65E3	
巔巔	T	彥彥	T	拋拋	T	昂昂	T
5DD3 5DD4		5F65 5F66		629B 62CB		6602 663B	
𨮒𨮒	T	德德	T	拔拔	TJ	晚晚	T
5E21 5E32		5FB3 5FB7		629C 62D4		665A 6669	
帶帶	TJ	徵徵	T	掙掙	T	暨暨	T
5E2F 5E36		5FB4 5FB5		6329 635D		66A8 66C1	
并并	T	惠惠	TJ	挿挿挿	TJ	曾曾	J
5E76 5E77		6075 60E0		633F 63D2 63F7		66FD 66FE	
廕廕	T	悅悅	T	捏捏	TJ	楞楞	T
5EC4 5ECF		6085 60A6		634F 63D1		67B4 67FA	
弑弑	T	悞悞	T	搜搜	TJ	查查	T
5F11 5F12		609E 60AE		635C 641C		67E5 67FB	
強強	T	𡇗𡇗	T	揭揭	T	柵柵	T
5F37 5F3A		60B3 60EA		63B2 63ED		67F5 6805	
彈彈	T	愠愠	T	搖搖搖	TJ	稅稅	T
5F39 5F3E		6120 614D		63FA 6416 6447		68B2 68C1	

榆榆 6961 6986	T	氫氫 6C32 6C33	T	滾滾 6EDA 6EFE	T	真真 771E 771F	TJ
概概 6982 69EA	T	汚汚 6C5A 6C61	T	潛潛 6F5B 6FF3	GTJK	眾眾 773E 8846	TJK
榼榼 6985 69B2	T	沒沒 6C92 6CA1	TJ	瀨瀨 7028 702C	T	研研 7814 784F	T
櫟櫟 699D 6A27	T	淨淨 6D44 6DE8	TJ	為爲 70BA 7232	GTJ	祿祿 797F 7984	TJ
楨楨 69C7 69D9	J	涉涉 6D89 6E09	T	煢煢 712D 7162	GTJK	禿禿 79BF 79C3	T
樣樣 69D8 6A23	TJ	況況 6D97 6D9A	T	熙熙 7155 7199	J	稅稅 7A05 7A0E	T
橫橫 6A2A 6A6B	T	淚淚 6D99 6DDA	T	煢煢 7174 7185	T	穗穗 7A42 7A57	TJ
步步 6B65 6B69	T	淥淥 6DE5 6E0C	T	狀狀 72B6 72C0	GT	箏箏 7B5D 7B8F	GJ
歲歲 6B72 6B73	T	清清 6DF8 6E05	T	瑤瑤 7464 7476	TJ	簞簞 7BB3 7C08	T
歿歿 6B7F 6B81	T	渴渴 6E07 6E34	T	瓶瓶 74F6 7501	T	篡篡 7BE1 7C12	T
殼殼 6BBB 6BBC	GTJ	溫溫 6E29 6EAB	T	產產 7522 7523	T	粵粵 7CA4 7CB5	T
毀毀 6BC0 6BC1	T	滄滄 6E88 6F59	T	瘦瘦 75E9 7626	J	絕絕 7D55 7D76	T
每每 6BCE 6BCF	T	漑漑 6E89 6F11	T	皞皞 76A1 76A5	T	綠綠 7DA0 7DD1	T

緒緒 7DD2 7DD6	T	菑菑 83D1 8458	TJ	說說 8AAA 8AAC	T	鄉鄉鄉 90F7 9109 9115	T
緣緣 7DE3 7E01	T	藎藎 8480 8495	T	諫諫 8ACC 8AEB	TJ	醞醞 9196 919E	T
緼緼 7DFC 7E15	T	蔣蔣 848B 8523	GJ	謠謠 8B20 8B21	J	醬醬 91A4 91AC	J
緼緼 7E48 7E66	T	薦薦 848D 853F	T	𢶏𢶏 8C5C 8C63	T	鉸鉸 9203 9292	T
羹羹 7FAE 7FB9	TJ	蒹蒹 8570 8580	T	走𠂔 8D70 8D71	TJ	銳銳 92B3 92ED	T
翺翺 7FF6 7FFA	T	薰薰 85AB 85B0	T	𢶏𢶏 8EFF 8F27	T	錄錄 9304 9332	T
𦍋𦍋 80FC 8141	T	蘊蘊 85F4 860A	T	輜輜 8F1C 8F3A	J	鍊鍊 932C 934A	TK
脫脫 812B 8131	T	虛虛 865A 865B	T	輶輶 8F3C 8F40	T	鎮鎮 93AD 93AE	TJ
膾膾 817D 8183	T	蛻蛻 86FB 8715	T	达达 8FBE 8FD6	T	閱閱 95B1 95B2	T
𦍋𦍋 8203 8204	GT	衛衛 885B 885E	TJK	迸迸 8FF8 902C	TJ	隄隄 9667 9689	G
舍舍 820D 820E	TJ	袞袞 886E 889E	TK	遙遙 9059 9065	J	青青 9751 9752	T
舖舖 8216 8217	J	裝裝 88C5 88DD	GJK	邢邢 90A2 90C9	T	靜靜 9759 975C	GTJ
莊莊 8358 838A	TJ	𦍋𦍋 8A2E 8A7D	T	郎郎 90CE 90DE	T	𢆶𢆶 976D 9771	J

頹頹 9839 983D	T	馱馱 99B1 99C4	TJK	鬪鬪 9B2C 9B2D	T	麪麪 9EAA 9EAB	T
顏顏 984F 9854	TJ	駢駢 99E2 9A08	TK	鯁鯁 9C1B 9C2E	TJ	麼麼 9EBC 9EBD	T
顛顛 985A 985B	J	飢飢 9AA9 9AAB	T	鳳鳳 9CEF 9CF3	T	黃黃 9EC3 9EC4	T
飲飲 98EE 98F2	J	高高 9AD8 9AD9	T	鶉鶉 9D87 9DAB	J	黑黑 9ED1 9ED2	T
餅餅 9905 9920	TJ	髮髮 9AEA 9AEE	TJ	鷓鷓 9DC6 9DCF	J		

S.4 Non-unification examples

In accordance with the unification procedures described in S.1 the pairs (or triplets) of ideographs shown below are not unified. The reason for non-unification is indicated by the reference which appears to the right of each pair (or triplet). For “non-cognate” see S.1.1.

NOTE – The reason for non-unification in these examples is different from the source separation rule described in S.1.6.

胃胃 5191 80C4	non cognate	寶寶 5BF3 5BF6	S.1.4.3	胸胸 6710 80CA	non cognate	稻稻 7A32 7A3B	S.1.4.3
冲冲 51B2 6C96	S.1.4.3	廳廳 5EF0 5EF3	S.1.4.1	眇眇 6713 8101	non cognate	翱翱 7FF1 7FF6	S.1.4.3
決決 51B3 6C7A	S.1.4.3	懷懷 61D0 61F7	S.1.4.1	腓腓 6718 8127	non cognate	耆耆耆 8007 8008 8009	S.1.4.3
況況 51B5 6CC1	S.1.4.3	𪗇𪗇 6560 656A	S.1.4.3	臃臃 6723 81A7	non cognate	聽聽聽 8074 807C 807D	S.1.4.1
堞堞 579B 579C	S.1.4.3	盼盼 670C 80A6	non cognate	朶朶 6735 6736	S.1.4.3	荊荊 8346 834A	S.1.4.2
孳孳 5B7C 5B7D	S.1.4.2	𦞙𦞙 670F 80D0	non cognate	灩灩 7054 7067	S.1.4.3	躲躲 8EB1 8EB2	S.1.4.3

Annex T
(informative)
Language tagging using Tag Characters

NOTE – Moved to F.8.

Annex U
(informative)
Characters in identifiers

A common task facing an implementer of UCS is the provision of a parsing and/or lexing engine for identifiers. Each programming language standard has its own identifier syntax; different programming languages have different conventions for the use of certain characters from the ASCII (ISO 646-IRV) range (\$, @, #, _) in identifiers. Questions as to which characters to use for syntactic purposes versus which to be allowed in identifiers, whether case-pairing should be included, normalization should be performed, and other factors enter into the picture when defining the set of permitted characters for a given identification purpose.

The Unicode Consortium publishes a document "UAX 31 – Identifier and Pattern Syntax" to assist in the standard treatment of identifiers in UCS character-based parsers. Those specifications are recommended for determining the list of UCS characters suitable for use in identifiers. The document is available at <http://www.unicode.org/reports/tr31/>.

