WG2 N3593

L2/09-162

ISO/IEC International Standard ISO/IEC 10646

Working Draft

Information technology – Universal Coded Character Set (UCS)

Technologie de l'information – Jeu universel de caractères codés (JUC)

Second edition, 2009

PDF disclaimer

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

Adobe is a trademark of Adobe Systems Incorporated.

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

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.ch
Web www.iso.ch

Printed in Switzerland

CONTENTS

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

14 Block and collection names 28
14.1 Block names 28
14.2 Collection names 29
15 Mirrored characters in bidirectional context 29
15.1 Mirrored characters 29
15.2 Directionality of bidirectional text 29
16 Special characters 29
16.1 Space characters 29
16.2 Currency symbols 30
16.3 Format Characters 30
16.4 Ideographic description characters 30
16.5 Variation selectors and variation sequences 31
17 Presentation forms of characters 33
18 Compatibility characters 34
19 Order of characters 34
20 Combining characters 34
20.1 Order of combining characters 34
20.2 Combining class and canonical ordering 34
20.3 Appearance in code charts 35
20.4 Alternate coded representations 35
20.5 Multiple combining characters 35
20.6 Collections containing combining characters 36
20.7 Combining Grapheme Joiner 36
21 Normalization forms 36
22 Special features of individual scripts and symbol repertoires 37
22.1 Hangul syllable composition method 37
22.2 Features of scripts used in India and some other South Asian countries 37
22.3 Byzantine musical symbols 37
23 Source references for CJK Ideographs 37
23.1 Source references for CJK Unified Ideographs 38
23.2 Source reference presentation for CJK Unified Ideographs 40
23.3 Source references for CJK Compatibility Ideographs 42
24 Character names and annotations 43
24.1 Entity names 43
24.2 Name formation 44
24.3 Single name 44
24.4 Name uniqueness 44
24.5 Character names for CJK Ideographs 45
24.6 Character names for Tangut characters 45
24.7 Character names for Hangul syllables 45
25 Named UCS Sequence Identifiers 47
26 Structure of the Basic Multilingual Plane 48
27 Structure of the Supplementary Multilingual Plane for scripts and symbols (SMP) 50
28 Structure of the Supplementary Ideographic Plane (SIP) 52

29 Structure of the Supplementary Special-purpose Plane (SSP) 52
30 Code charts and lists of character names 52
30.1 Code chart 53
30.2 Character names list 53
30.3 Pointers to code charts and lists of character names 54
Annex A (normative) Collections of graphic characters for subsets 2151
A.1 Collections of coded graphic characters 2151
A.2 Blocks lists 2156
A.3 Fixed collections of the whole UCS (except Unicode collections) 2158
A.4 CJK collections 2161
A.5 Other collections 2162
A.6 Unicode collections 2165
Annex B (normative) List of combining characters 2173
Annex C (normative) Transformation format for planes 1 to 10 of the UCS (UTF-16) 2174
Annex D (normative) UCS Transformation Format 8 (UTF-8) 2175
Annex E (normative) Mirrored characters in bidirectional context 2176
Annex F (informative) Format characters 2177
F.1 General format characters 2177
F.2 Script-specific format characters 2179
F.3 Interlinear annotation characters 2180
F.4 Subtending format characters 2180
F.5 Western musical symbols 2180
F.6 Language tagging using Tag characters 2181
Annex G (informative) Alphabetically sorted list of character names 2183
Annex H (informative) The use of “signatures” to identify UCS 2184
Annex I (informative) Ideographic description characters 2185
Annex J (informative) Recommendation for combined receiving/originating devices with internal storage 2188
Annex K (informative) Notations of octet value representations 2189
Annex L (informative) Character naming guidelines 2190
Annex M (informative) Sources of characters 2193
Annex N (informative) External references to character repertoires 2197
N.1 Methods of reference to character repertoires and their coding 2197
N.2 Identification of ASN.1 character abstract syntaxes 2197
N.3 Identification of ASN.1 character transfer syntaxes 2198
Annex P (informative) Additional information on CJK Unified Ideographs 2199
Annex Q (informative) Code mapping table for Hangul syllables 2200
Annex R (informative) Names of Hangul syllables 2201
Annex S (informative) Procedure for the unification and arrangement of CJK Ideographs 2202
S.1 Unification procedure 2202
S.2 Arrangement procedure 2205
S.3 Source code separation examples 2206
Annex T (informative) Language tagging using Tag Characters 2212

Annex U (informative) Characters in identifiers 2213

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 JTC1.

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

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

Attention is drawn to the possibility that some of the elements of ISO/IEC 10646 may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

International Standard ISO/IEC 10646 was prepared by Joint Technical Committee ISO/IEC JTC1, Information technology, Subcommittee SC 2, Coded Character sets.

This second edition of ISO/IEC 10646 cancels and replaces ISO/IEC 10646:2003. It also incorporates ISO/IEC 10646:2003 Amd.1:2005, Amd.2:2006, Amd.3:2007, Amd.4:2008, Amd.5:2009, Amd.6:2010, and Amd.7:2011.

NOTE – Amendment 6 and 7 are still in progress. The text in this document is synchronized with their contents and will be updated accordingly.

Introduction

ISO/IEC 10646 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. ISO/IEC 10646 has been widely adopted in new Internet protocols and implemented in modern operating systems and computer languages. This edition covers over 99 000 characters from the world's scripts.

ISO/IEC 10646 contains material which may only be available to users who obtain their copy in a machine readable format. That material consists of the following printable files:

- CJKU_SR.txt
- CJKC_SR.txt
- IICORE.txt
- JIEx.txt
- Allnames.txt
- HangulSy.txt.

Information technology — Universal Coded Character Set (UCS) —

1 Scope

ISO/IEC 10646 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 document

- specifies the architecture of ISO/IEC 10646,
- defines terms used in ISO/IEC 10646,
- 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 a encoding system different from that specified in ISO/IEC 2022. The method to designate UCS from ISO/IEC 2022 is specified in 12.2.

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

> NOTE – The Unicode Standard, Version 5.2 includes a set of characters, names, and coded representations that are identical with those in this International Standard. It additionally provides details of character properties, processing algorithms, and definitions that are useful to implementers.

2 Conformance

2.1 General

Whenever private use characters are used as specified in ISO/IEC 10646, the characters themselves shall not be covered by these conformance requirements.

2.2 Conformance of information interchange

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

a) all the coded representations of graphic characters within that CC-data-element conform to 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 CC-data-element are taken from those within an identified subset (see 8);

c) all the coded representations of control functions within that CC-data-element 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 ISO/IEC 10646 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 ISO/IEC 10646 shall be the subject of a description that identifies the means by which the user may supply characters to the device and/or may recognize them when they are made available to the user, as specified respectively, in subclauses b) and c) below.

b) **Originating device**: An originating device shall allow its user to supply any characters from an adopted subset, and be capable of transmitting their coded representations within a CC-data-element in accordance with the adopted encoding form and adopted encoding scheme. As such, the originating device shall not emit ill-formed CC-data-elements.

c) **Receiving device**: A receiving device shall be capable of receiving and interpreting any coded representation of characters that are within a CC-data-element in accordance with the adopted 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 CC-data-elements 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 standard.

NOTE 2 – See also 0Annex J for receiving devices with retransmission capability.

3 Normative references

The following normative documents contain provisions which, through reference in this text, constitute provisions of ISO/IEC 10646. For dated references, subsequent amendments to, or revisions of, any of these publications do not apply. However, parties to agreements based on ISO/IEC 10646 are encouraged to investigate the possibility of applying the most recent editions of the normative documents indicated below. For undated references, the latest edition of the normative document referred to applies. Members of ISO and IEC maintain registers of currently valid International Standards.

ISO/IEC 2022:1994 *Information technology — Character code structure and extension techniques.*

ISO/IEC 6429:1992 *Information technology — Control functions for coded character sets.*

Unicode Standard Annex, UAX#9, *The Unicode Bidirectional Algorithm, Version 6.0.0, [Date TBD].*

Unicode Standard Annex, *UAX#15, Unicode Normalization Forms, Version 6.0.0, [Date TBD].*

Unicode Standard Annex, *UAX#34, Unicode Named Character Sequences, Version 6.0, [Date TBD].*

Unicode Technical Standard, *UTS#37, Ideographic Variation Database, Version 1.0, January 2006.*

Unicode Standard Annex, *UAX#44, Unicode Character Database Version 6.2, [Date TBD].*

4 Terms and definitions

For the purposes of ISO/IEC 10646, the following terms and definitions apply.

4.1
Base character
A graphic character which is not a combining character

NOTE – 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

4.2
Basic Multilingual Plane
BMP
Plane 00 of the UCS codespace

4.3
Block
A 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 representation
The representation with which characters of this coded character set are specified using code points within the UCS codespace

4.5
CC-data-element
coded-character-data-element
code unit sequence
An 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; such sequence may contain code units associated with any type of code points

NOTE – Unlike previous editions of the standard, this version does not use anymore implementation levels. Its definition of CC-data-element content corresponds to the former unrestricted implementation level 3. Other definitions of CC-data-element content, 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 standards such as ISO/IEC 8824 and ISO/IEC 8825, the concept of implementation level may still be referenced as 'Implementation level 3'. See Annex N.

4.6
Character
A member of a set of elements used for the organization, control, or representation of textual data; a character may be represented by a sequence of one or several coded characters

4.7
Character boundary
Within a CC-data-element the demarcation between the last code unit of a coded character and the first code unit of the next coded character

4.8
Code chart
Code table
A rectangular array showing the representation of coded characters allocated within a range of the UCS codespace

4.9
Coded character
An association between a character and a code point

4.10
Coded character set
A set of coded characters

4.11
Code point
Code position
Any value in the UCS codespace; the term code point is preferred

4.12
Code unit
The 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 unit) 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.

4.13
Collection
A numbered and named set of entities; 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 4.24 for extended collection)

NOTE – 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
Characters which have General Category values of Spacing Combining Mark (Mc), Non Spacing Mark (Mn), and Enclosing Mark (Me) according to the Unicode Character Database (see 3).

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
A graphic character included as a coded character of ISO/IEC 10646 primarily for compatibility with existing coded character sets

4.17
Composite sequence
A 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 may be used to represent characters not encoded in the repertoire of ISO/IEC 10646

4.18
Control character
A control function the coded representation of which represents 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 11 for a list of the long names used by ISO/IEC 6429 in association with the control characters.

4.19
Control function
An action that affects the recording, processing, transmission, or interpretation of data, and that is represented by a CC-data-element

4.20
Default state
The state that is assumed when no state has been explicitly specified (see F.2.1 and F.2.2)

4.21
Device
A component of information processing equipment which can transmit and/or receive coded information within CC-data-elements (It may be an input/output device in the conventional sense, or a process such as an application program or gateway function.)

4.22
Encoding form
An encoding form determines how each UCS code point for a UCS character is to be expressed as one or more code unit used by the encoding form. ISO/IEC 10646 specifies UTF-8, UTF-16, and UTF-32

4.23
Encoding scheme
An encoding scheme 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 context. UCS encoding forms refer to in-memory and application interface representation of textual data. UCS encoding schemes refer to octet-serialized textual data.

4.24
Extended collection
A collection for which the entities can also consist of sequences of code points that are in normalization form NFC (see 21); the sequences of code points are referenced by Named UCS Sequence Identifiers (NUSI) (see 125).

NOTE – 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

4.25
Fixed collection
A 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

4.26
Format character
A character whose primary function is to affect the layout or processing of characters around it; it generally does not have a visible representation of its own

4.27
General Category
GC
Value assigned to each UCS code point which determines its major class, such as letter, punctuation, and symbol; each value is defined as a two-letter abbreviation in the Unicode Character Database (see 3)

NOTE – 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'.

4.28
Graphic character
A character, other than a control function or a format character, that has a visual representation normally handwritten, printed, or displayed

4.29
Graphic symbol
The visual representation of a graphic character or of a composite sequence

4.30
High-surrogate code point
A code point in the range D800 to DBFF reserved for the use of UTF-16

4.31
High-surrogate code unit
A 16-bit code unit in the range D800 to DBFF used in UTF-16 as the leading code unit of a surrogate pair (see 9.2)

4.32
ill-formed CC-data-element
A UCS CC-data-element that purports to be in a UCS encoding form which does not conform to the specification of that encoding form (for example, an unpaired surrogate code unit is an ill-formed CC-data-element)

4.33
ill-formed CC-data-element subset
A non-empty subset of a CC-data-element X which does not contain any code unit which also belong to any minimal well-formed CC-data-element subset of X

NOTE – An ill-formed CC-data-element subset cannot overlap with a minimal well-formed CC-data-element.

4.34
Interchange
The transfer of character coded data from one user to another, using telecommunication means or interchangeable media; interchange implies data serialization and the usage of a UCS encoding scheme

4.35
Interworking
The process of permitting two or more systems, each employing different coded character sets, meaningfully to interchange character coded data; conversion between the two codes may be involved

4.36
ISO/IEC 10646-1
A former subdivision of the standard. It is also referred to as Part 1 of ISO/IEC 10646 and contained the specification of the overall architecture and the Basic Multilingual Plane (BMP). There are a First and a Second Edition of ISO/IEC 10646-1.

4.37
ISO/IEC 10646-2
A former subdivision of the standard. It is also referred to as Part 2 of ISO/IEC 10646 and contained the specification of the Supplementary Multilingual Plane (SMP), the Supplementary Ideographic Plane (SIP) and the Supplementary Special-purpose Plane (SSP). There is only a First Edition of ISO/IEC 10646-2.

4.38
Low-surrogate code point
A code point in the range DC00 to DFFF reserved for the use of UTF-16

4.39
Low-surrogate code unit
A 16-bit code unit in the range DC00 to DFFF used in UTF-16 as the trailing code unit of a surrogate pair (see 9.2)

4.40
Minimal well-formed CC-data-element
A well-formed CC-data-element that maps to a single UCS scalar value

4.41
Mirrored character
A character whose image is mirrored horizontally in text that is laid out from right to left

4.42
Octet
A 8-bit code unit; the value is expressed in hexadecimal notation from 00 to FF in ISO/IEC 10646 (see Annex K)

4.43
Plane
A subdivision of the UCS codespace consisting of 65536 code points. The UCS codespace contain 17 planes.

4.44
Presentation;
to present
The process of writing, printing, or displaying a graphic symbol.

4.45
Presentation form
In the presentation of some scripts, a form of a graphic symbol representing a character that depends on the position of the character relative to other characters

4.46
Private use plane
A plane within this coded character set; the contents of which is not specified in ISO/IEC 10646. Planes 0F and 10 are private use planes

4.47
Repertoire
A specified set of characters that are represented in a coded character set

4.48
Row
A subdivision of a plane; by multiple of 256 code points

4.49
Script
A set of graphic characters used for the written form of one or more languages

4.50
Supplementary plane
A plane other than Plane 00 of the UCS codespace; a plane that accommodates characters which have not been allocated to the Basic Multilingual Plane

4.51
Supplementary Multilingual Plane for scripts and symbols
SMP
Plane 01 of the UCS codespace

4.52
Supplementary Ideographic Plane
SIP
Plane 02 of the UCS codespace

4.53
Supplementary Special-purpose Plane
SSP
Plane 0E of the UCS codespace

4.54
Surrogate pair
A 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

4.55
Tertiary Ideographic Plane
TIP
Plane 03 of the UCS codespace

4.56
UCS codespace
The UCS codespace consists of the integers from 0 to 10FFFF (hexadecimal) available for assigning the repertoire of the UCS characters

4.57
UCS scalar value
Any UCS code point except high-surrogate and low-surrogate code points

4.58
Unpaired surrogate code unit
A surrogate code unit in a CC-data element 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

4.59
User
A person or other entity that invokes the service provided by a device (This entity may be a process such as an application program if the “device” is a code converter or a gateway function, for example.)

4.60
Well-formed CC-data-element
A UCS CC-data-element that purports to be in a UCS encoding form which conforms to the specification of that encoding form and contains no ill-formed CC-data-element 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.

ISO/IEC 10646 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 ISO/IEC 10646 shall be regarded as a single entity made of 17 planes.

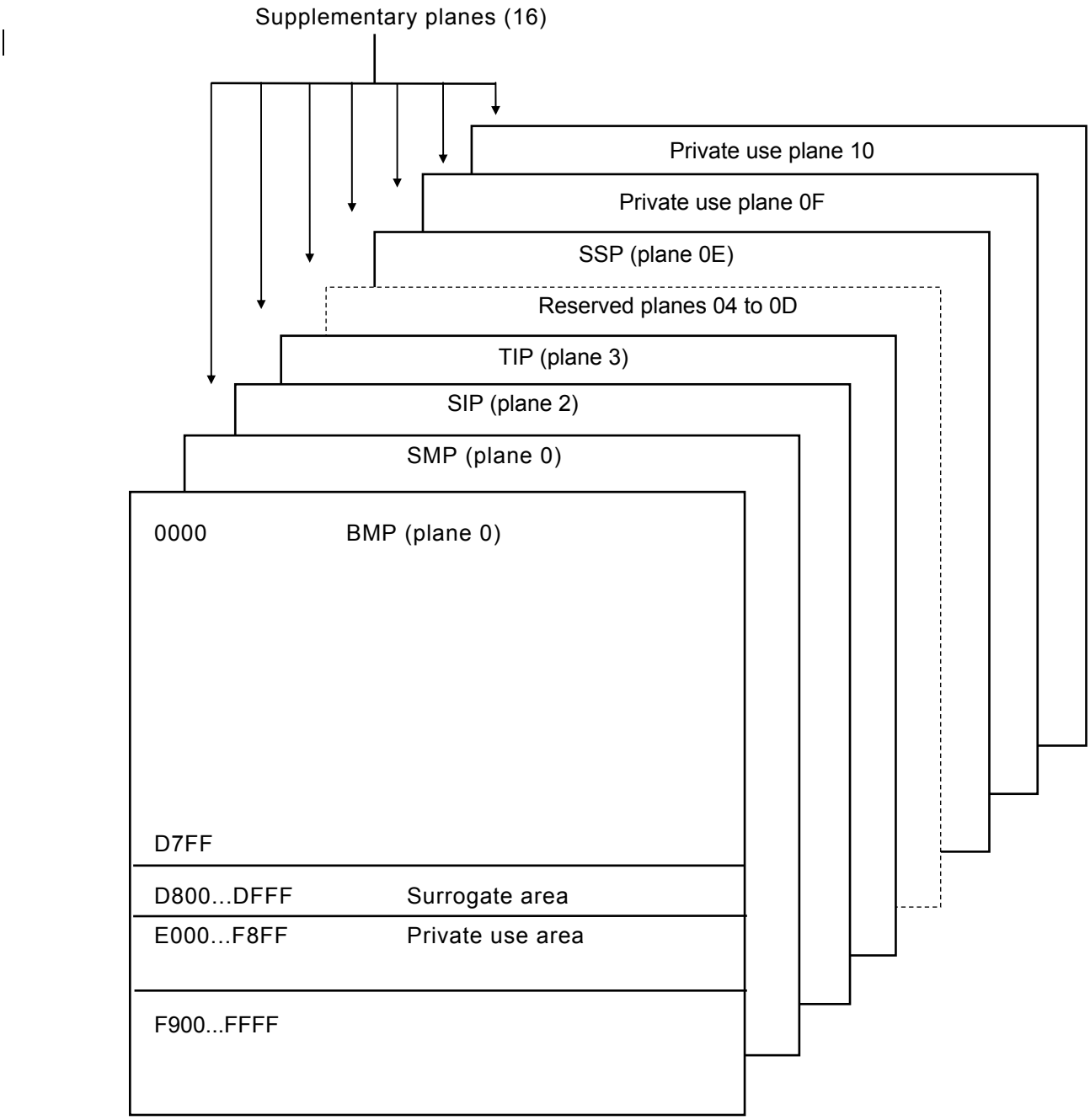


Figure 1 - Planes of the Universal Coded Character Set

6.2 Coding of characters

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

When a single character is to be identified in term 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 Type of code points

6.3.1 Classification

UCS code points are categorized in basic types, according to their General Category value. 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 purpose and have different character names.

6.3.3 Format characters

Code points 2060 to 206F, FFF0 to FFFC, and E0000 to E0FFF are reserved for Format Characters (see 16.3 and Annex F).

NOTE – Unassigned code points in those ranges may be ignored in normal processing and display.

6.3.4 Control characters

Code points 0000 to 001F, 007F to 009F in the BMP are reserved for control characters (see 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 ISO/IEC 10646. 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 ISO/IEC 10646, 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 9.2). 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 ISO/IEC 10646 will not allocate any characters to code points reserved for private use characters or for transformation formats.

6.4 Naming of characters

ISO/IEC 10646 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 24.5 for Chinese /Japanese/Korean (CJK) ideographs, or

d) follows the rule given in 24.6 for Tangut characters, or

e) follows the rule given in 24.7 for Hangul syllables.

Additional rules to be used for constructing the names of characters are given in 24.2.

The list of character names except for CJK ideographs, Hangul syllables, and Tangut characters is provided by 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)

ISO/IEC 10646 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 that, 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 | u } [{+}(xxxx | xxxxx | 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

ISO/IEC 10646 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:

<UID1, UID2, ..., UIDn>

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.

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 10), ISO/IEC 10646 defines an identifier for serialized octet sequence. For a sequence of n octets it has the following form:

<xx_1 xx_2 … xx_n>

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.

NOTE – It is intended that in future editions of ISO/IEC 10646, the names and allocation of the characters in this edition will remain unchanged.

8 Subsets

ISO/IEC 10646 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.

8.1 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 inter-work 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 ISO/IEC 10646.

8.2 Selected subset

A selected subset consists of a list of collections of graphic characters as defined in ISO/IEC 10646. 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 ISO/IEC 10646.

9 UCS encoding forms

ISO/IEC 10646 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.

9.1 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 CC-sequences for these octet values.
- The first octet in the UTF-8 coded representation of any character can be directly identified when a CC-data-element 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	1st octet	2nd octet	3rd octet	4th octet
00000000 0xxxxxxx	0xxxxxxx			
00000yyy yyxxxxxx	110yyyyy	10xxxxxx		
zzzzyyyy yyxxxxx	1110zzzz	10yyyyyy	10xxxxxx	
000uuuuu zzzzyyyy yyxxxxxx	11110uuu	10uuzzzz	10yyyyyy	10xxxxxx

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	1st octet	2nd octet	3rd octet	4th 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-FFFFF	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

9.2 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 to account for the starting offset to the scalar value (expressed as 'wwww = uuuuu-1' in the table).

Table 4: UTF-16 Bit distribution

Scalar value	UTF-16
xxxxxxxxxxxxxxxx	xxxxxxxxxxxxxxxx
000uuuuuxxxxxxxxxxxxxxxx	110110wwwwxxxxxx 110111xxxxxxxxxx

NOTE – Former editions of this 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.

9.3 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

Encoding schemes are octet serialization 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.

ISO/IEC 10646 specifies seven encoding schemes: UTF-8, UTF-16BE, UTF-16LE, UTF-16, UTF-32BE, UTF-32LE, and UTF-32.

10.1 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

10.2 UTF-16BE.

The UTF-16BE encoding scheme serializes a UTF-16 CC-data-element 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.

10.3 UTF-16LE

The UTF-16LE encoding scheme serializes a UTF-16 CC-data-element 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.

10.4 UTF-16

The UTF-16 encoding scheme serializes a UTF-16 CC-data-element 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.

10.5 UTF-32BE

The UTF-32BE encoding scheme serializes a UTF-32 CC-data-element 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.

10.6 UTF-32LE

The UTF-32LE encoding scheme serializes a UTF-32 CC-data-element 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.

10.7 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 octet precedes or follows the more significant octet.

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 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 ISO/IEC 10646 the action of those control functions will depend on the type of element from ISO/IEC 10646 that has been chosen, by the application, to be the element (or character) on which the control functions act. These 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
0001 START OF HEADING
0002 START OF TEXT
0003 END OF TEXT

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

The control character 0084 INDEX has been removed from ISO/IEC 6492:1992. 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

CC-data-elements conforming to ISO/IEC 10646 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 ISO/IEC 10646 (including the encoding form and the encoding scheme) and any subset of the coding space that have been adopted by the originator must also be available to the recipient. The route by which such identification is communicated to the recipient is outside the scope of ISO/IEC 10646.

However, some standards for interchange of coded information may permit, or require, that the coded representation of the identification applicable to the CC-data-element forms a part of the interchanged information. This clause specifies a coded representation for the identification of UCS and a subset of ISO/IEC 10646, and also of a C0 and a C1 set of control functions from ISO/IEC 6429 for use in conjunction with ISO/IEC 10646. 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 this clause.

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

12.2 Identification of a UCS encoding form

When the escape sequences from ISO/IEC 2022 are used, the identification of a UCS encoding form (see 9) specified by ISO/IEC 10646 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 1 – 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, 02/05 02/15 04/07, 02/05 02/15 04/08, 02/05 02/15 04/10, 02/05 02/15 04/011 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 CC-data-element content definition.

NOTE 2 – The following escape sequence may also be used:

ESC 02/05 04/07
UTF-8 encoding form; UTF-8 encoding scheme

The escape sequence used for a return to the coding system of ISO/IEC 2022 is not padded (see 12.5).

If such an escape sequence appears within a CC-data-element 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 CC-data-element conforming to ISO/IEC 10646, it shall be padded in accordance with clause 11.

12.3 Identification of subsets of graphic characters

When the control sequences of ISO/IEC 6429 are used, the identification of subsets (see 8) specified by ISO/IEC 10646 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 Annex A of ISO/IEC 10646. 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 CC-data-element 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 CC-data-element conforming to ISO/IEC 10646, 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

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 CC-data-element 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 CC-data-element conforming to ISO/IEC 10646, 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 CC-data-element conforming to ISO/IEC 10646, it shall be padded in accordance with clause 11.

If such an escape sequence appears within a CC-data-element 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 in clause 12.2 for the identification of UCS 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 30 sets out the detailed code charts and the lists of character names for the graphic characters. It specifies graphic characters, their coded representation, and the character name for each character.

NOTE – Clause 30 also includes additional information on characters clarifying some feature of a character, such as its naming or usage, or its associated graphic symbol.

The graphic symbols are to be regarded as typical visual representations of the characters. ISO/IEC 10646 does not attempt to prescribe the exact shape of each character. The shape is affected by the design of the font employed, which is outside the scope of ISO/IEC 10646.

Graphic characters specified in ISO/IEC 10646 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 ISO/IEC 10646; 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 chart.

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 3 to 7, 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 2 to 6.

Rules to be used for constructing the names of blocks are given in 24.4.1.

14.2 Collection names

Collections are shown in Annex A.

Rules to be used for constructing the names of collections are given in 24.4.2.

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 CC-data-element 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 Character Database (see 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 Character Database (see 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 3) describes the algorithm used to determine the directionality for bidirectional text.

16 Special characters

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

16.1 Space characters

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

Code Point	Name	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
180E	MONGOLIAN VOWEL SEPARATOR	2008	PUNCTUATION SPACE
2000	EN QUAD	2009	THIN SPACE
2001	EM QUAD	200A	HAIR SPACE
2002	EN SPACE	202F	NARROW NO-BREAK SPACE
2003	EM SPACE	205F	MEDIUM MATHEMATICAL SPACE
2004	THREE-PER-EM SPACE	3000	IDEOGRAPHIC SPACE

NOTE 1 – The character 180E MONGOLIAN VOWEL SEPARATOR may be used between the MONGOLIAN LETTER A or the MONGOLIAN LETTER E at the end of a word and the preceding consonant letter. It indicates a special form of the graphic

symbol for the letter A or E and the preceding consonant. When rendered in visible form it is generally shown as a narrow space between the letters, but it may sometimes be shown as a distinct graphic symbol to assist the user.

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.

16.2 Currency symbols

Currency symbols in ISO/IEC 10646 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.

16.3 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
0600	ARABIC NUMBER SIGN
0601	ARABIC SIGN SANAH
0602	ARABIC FOOTNOTE MARKER
0603	ARABIC SIGN SAFHA
06DD	ARABIC END OF AYAH
070F	SYRIAC ABBREVIATION MARK
17B4	KHMER VOWEL INHERENT AQ
17B5	KHMER VOWEL INHERENT AA
1A60	TAI THAM SIGN SAKOT
1CBF	MEITEI MAYEK SIGN VIRAMA
200B	ZERO WIDTH SPACE
200C	ZERO WIDTH NON-JOINER
200D	ZERO WIDTH JOINER
200E	LEFT-TO-RIGHT MARK
200F	RIGHT-TO-LEFT MARK
2028	LINE SEPARATOR
2029	PARAGRAPH SEPARATOR
202A	LEFT-TO-RIGHT EMBEDDING
202B	RIGHT-TO-LEFT EMBEDDING
202C	POP DIRECTIONAL FORMATTING
202D	LEFT-TO-RIGHT OVERRIDE
202E	RIGHT-TO-LEFT OVERRIDE
2060	WORD JOINER
2061	FUNCTION APPLICATION
2062	INVISIBLE TIMES
2063	INVISIBLE SEPARATOR
2064	INVISIBLE PLUS
206A	INHIBIT SYMMETRIC SWAPPING
206B	ACTIVATE SYMMETRIC SWAPPING
206C	INHIBIT ARABIC FORM SHAPING
206D	ACTIVATE ARABIC FORM SHAPING
206E	NATIONAL DIGIT SHAPES
206F	NOMINAL DIGIT SHAPES
2D7F	TIFINAGH CONSONANT JOINER
FEFF	ZERO WIDTH NO-BREAK SPACE
FFF9	INTERLINEAR ANNOTATION ANCHOR
FFFA	INTERLINEAR ANNOTATION SEPARATOR
FFFB	INTERLINEAR ANNOTATION TERMINATOR
110BD	KAITHI NUMBER SIGN
1D173	MUSICAL SYMBOL BEGIN BEAM
1D174	MUSICAL SYMBOL END BEAM
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
E0001	LANGUAGE TAG
E0020-E007F	TAG SPACE to CANCEL TAG

16.4 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

16.5 Variation selectors and variation sequences

Variation selectors are a specific class of combining characters immediately following a non decomposable base character and which indicate a specific variant form of graphic symbol for that character. A decomposable character is a character for which there exists an equivalent composite sequence. The character sequence consisting of a non decomposable base character followed by a variation selector is called a variation sequence.

NOTE 1 – Some variation selectors are specific to a script, such as the Mongolian free variation selectors, others are used with various other base characters such as the mathematical symbols.

Only the variation sequences defined or referenced in this clause 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.

No variation sequences using characters from VARIATION SELECTOR-2 to VARIATION SELECTOR-16 are defined at this time. 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 (see 3).

NOTE 2 – This version of the standard incorporates by reference the variation sequences listed in version 2007-12-14 of the Ideographic Variation Database, as described at http://www.unicode.org/ivd/data/2007-12-14/.

The following list provides a description of the variant appearances corresponding to the use of appropriate variation selectors with all allowed base mathematical symbols.

NOTE 3 – The VARIATION SELECTOR-1 is the only variation selector used with mathematical symbols.

Sequence (UID notation)	Description of variant appearance
<2229, FE00>	INTERSECTION with serifs
<222A, FE00>	UNION with serifs
<2268, FE00>	LESS-THAN BUT NOT EQUAL TO with vertical stroke
<2269, FE00>	GREATER-THAN BUT NOT EQUAL TO with vertical stroke
<2272, FE00>	LESS-THAN OR EQUIVALENT TO following the slant of the lower leg
<2273, FE00>	GREATER-THAN OR EQUIVALENT TO following the slant of the lower leg
<228A, FE00>	SUBSET OF WITH NOT EQUAL TO with stroke through bottom members
<228B, FE00>	SUPERSET OF WITH NOT EQUAL TO with stroke through bottom members
<2293, FE00>	SQUARE CAP with serifs
<2294, FE00>	SQUARE CUP with serifs
<2295, FE00>	CIRCLED PLUS with white rim
<2297, FE00>	CIRCLED TIMES with white rim
<229C, FE00>	CIRCLED EQUALS equal sign touching the circle
<22DA, FE00>	LESS-THAN EQUAL TO OR GREATER-THAN with slanted equal
<22DB, FE00>	GREATER-THAN EQUAL TO OR LESS-THAN with slanted equal
<2A3C, FE00>	INTERIOR PRODUCT tall variant with narrow foot
<2A3D, FE00>	RIGHTHAND INTERIOR PRODUCT tall variant with narrow foot
<2A9D, FE00>	SIMILAR OR LESS-THAN with similar following the slant of the upper leg
<2A9E, FE00>	SIMILAR OR GREATER-THAN with similar following the slant of the upper leg
<2AAC, FE00>	SMALLER THAN OR EQUAL TO with slanted equal
<2AAD, FE00>	LARGER THAN OR EQUAL TO with slanted equal
<2ACB, FE00>	SUBSET OF ABOVE NOT EQUAL TO with stroke through bottom members
<2ACC, FE00>	SUPERSET OF ABOVE NOT EQUAL TO with stroke through bottom members

The following list provides a description of the variant appearances corresponding to the use of appropriate variation selectors with all allowed base Mongolian characters. Only some presentation forms of the base Mongolian characters used with the Mongolian free variation selectors produce variant appearances.

NOTE 4 – The Mongolian characters have various presentation forms depending on their position in a CC-data element. These presentations forms are called isolate, initial, medial and final.

Sequence (UID notation)	position	Description of variant appearance
<1820, 180B>	isolate, medial, final	MONGOLIAN LETTER A second form
<1820, 180C>	medial	MONGOLIAN LETTER A third form
<1821, 180B>	initial, final	MONGOLIAN LETTER E second form
<1822, 180B>	medial	MONGOLIAN LETTER I second form
<1823, 180B>	medial,final	MONGOLIAN LETTER O second form
<1824, 180B>	medial	MONGOLIAN LETTER U second form
<1825, 180B>	medial,final	MONGOLIAN LETTER OE second form
<1825, 180C>	medial	MONGOLIAN LETTER OE third form
<1826, 180B>	isolate, medial, final	MONGOLIAN LETTER UE second form
<1826, 180C>	medial	MONGOLIAN LETTER UE third form
<1828, 180B>	initial, medial	MONGOLIAN LETTER NA second form
<1828, 180C>	medial	MONGOLIAN LETTER NA third form
<1828, 180D>	medial	MONGOLIAN LETTER NA separate form
<182A, 180B>	final	MONGOLIAN LETTER BA alternative form
<182C, 180B>	initial, medial	MONGOLIAN LETTER QA second form
<182C, 180B>	isolate	MONGOLIAN LETTER QA feminine second form
<182C, 180C>	medial	MONGOLIAN LETTER QA third form
<182C, 180D>	medial	MONGOLIAN LETTER QA fourth form
<182D, 180B>	initial, medial	MONGOLIAN LETTER GA second form
<182D, 180B>	final	MONGOLIAN LETTER GA feminine form
<182D, 180C>	medial	MONGOLIAN LETTER GA third form
<182D, 180D>	medial	MONGOLIAN LETTER GA feminine form
<1830, 180B>	final	MONGOLIAN LETTER SA second form
<1830, 180C>	final	MONGOLIAN LETTER SA third form
<1832, 180B>	medial	MONGOLIAN LETTER TA second form
<1833, 180B>	initial, medial, final	MONGOLIAN LETTER DA second form
<1835, 180B>	final	MONGOLIAN LETTER JA second form
<1836, 180B>	initial, medial	MONGOLIAN LETTER YA second form
<1836, 180C>	medial	MONGOLIAN LETTER YA third form
<1838, 180B>	final	MONGOLIAN LETTER WA second form
<1844, 180B>	medial	MONGOLIAN LETTER TODO E second form
<1845, 180B>	medial	MONGOLIAN LETTER TODO I second form
<1846, 180B>	medial	MONGOLIAN LETTER TODO O second form
<1847, 180B>	isolate, medial, final	MONGOLIAN LETTER TODO U second form
<1847, 180C>	medial	MONGOLIAN LETTER TODO U third form
<1848, 180B>	medial	MONGOLIAN LETTER TODO OE second form
<1849, 180B>	isolate, medial	MONGOLIAN LETTER TODO UE second form
<184D, 180B>	initial, medial	MONGOLIAN LETTER TODO QA feminine form
<184E, 180B>	medial	MONGOLIAN LETTER TODO GA second form
<185D, 180B>	medial, final	MONGOLIAN LETTER SIBE E second form
<185E, 180B>	medial, final	MONGOLIAN LETTER SIBE I second form

<185E, 180C>	medial, final	MONGOLIAN LETTER SIBE I third form
<1860, 180B>	medial, final	MONGOLIAN LETTER SIBE UE second form
<1863, 180B>	medial	MONGOLIAN LETTER SIBE KA second form
<1868, 180B>	initial, medial	MONGOLIAN LETTER SIBE TA second form
<1868, 180C>	medial	MONGOLIAN LETTER SIBE TA third form
<1869, 180B>	initial, medial	MONGOLIAN LETTER SIBE DA second form
<186F, 180B>	initial, medial	MONGOLIAN LETTER SIBE ZA second form
<1873, 180B>	medial, final	MONGOLIAN LETTER MANCHU I second form
<1873, 180C>	medial, final	MONGOLIAN LETTER MANCHU I third form
<1873, 180D>	medial	MONGOLIAN LETTER MANCHU I fourth form
<1874, 180B>	medial	MONGOLIAN LETTER MANCHU KA second form
<1874, 180B>	final	MONGOLIAN LETTER MANCHU KA feminine first form
<1874, 180C>	medial	MONGOLIAN LETTER MANCHU KA feminine first form
<1874, 180C>	final	MONGOLIAN LETTER MANCHU KA feminine second form
<1874, 180D>	medial	MONGOLIAN LETTER MANCHU KA feminine second form
<1876, 180B>	initial, medial	MONGOLIAN LETTER MANCHU FA second form
<1880, 180B>	all	MONGOLIAN LETTER ALI GALI ANUSVARA ONE second form
<1881, 180B>	all	MONGOLIAN LETTER ALI GALI VISARGA ONE second form
<1887, 180B>	isolate, final	MONGOLIAN LETTER ALI GALI A second form
<1887, 180C>	final	MONGOLIAN LETTER ALI GALI A third form
<1887, 180D>	final	MONGOLIAN LETTER ALI GALI A fourth form
<1888, 180B>	final	MONGOLIAN LETTER ALI GALI I second form
<188A, 180B>	initial, medial	MONGOLIAN LETTER ALI GALI NGA second form

The following list provides a description of the variant appearances corresponding to the use of appropriate variation selectors with all allowed base Phags-pa characters. These variation selector sequences do not select fixed visual representation; rather, they select a representation that is reversed from the normal form predicted by the preceding character.

Sequence (UID notation)	Description of variant appearance
<A856, FE00>	PHAGS-PA LETTER reversed shaping SMALL A
<A85C, FE00>	PHAGS-PA LETTER reversed shaping HA
<A85E, FE00>	PHAGS-PA LETTER reversed shaping I
<A85F, FE00>	PHAGS-PA LETTER reversed shaping U
<A860, FE00>	PHAGS-PA LETTER reversed shaping E
<A868, FE00>	PHAGS-PA SUBJOINED LETTER reversed shaping YA

NOTE 5 – The variation selector only selects a different *appearance* of an already encoded character. It is not intended as a general code extension mechanism.

NOTE 6 – The exhaustive list of standardized variants is also described as *StandardizedVariants.html* in the Unicode character database (http://www.unicode.org/Public/5.0.0/ucd/StandardizedVariants.html).

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 ISO/IEC 10646.

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 ISO/IEC 10646.

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 ISO/IEC 10646 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.

19 Order of characters

Usually, coded characters appear in a CC-data-element 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 CC-data-element.

20 Combining characters

This clause specifies the use of combining characters (see 4.14).

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 00AD NO-BREAK SPACE. For example, grave accent can be composed as 00AD NO-BREAK SPACE followed by 0300 COMBINING GRAVE ACCENT.

NOTE – Indic matras 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 Indic matra with the character SPACE.

20.2 Combining class and canonical ordering

Each combining character has a combining class value determined by the Unicode Character Database (see 3). The combining class is used to determine the canonical ordering which is part of the normalization process (see 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 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. ISO/IEC 10646 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 re-presentations 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 ISO/IEC 10646.

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 should be ignored unless it specifically occurs with a tailored collation element mapping. For rendering, the combining grapheme joiner is invisible.

NOTE 1 – 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 ISO/IEC 10646 are specified in the Unicode Standard UAX#15 (see 3). There are four normalization forms:

1) Normalization Form D (NFD) which is a canonical decomposition,
2) Normalization Form C (NFC) which is a canonical decomposition followed by canonical composition,
3) Normalization Form KD (NFKD) which is a compatibility decomposition,
4) Normalization Form KC (NFKC) which is a compatibility decomposition followed by canonical composition.

NOTE 1 – The result of applying any of these normalization forms onto a CC-data-element is intended to stay stable over time. It means that the normalized representation of a CC-data-element 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.6.2).

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

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) 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 or a Jongseong shall be preceded by a CHOSEONG FILLER (115F). 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.

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 30) 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 0B Tamil

The graphic symbol for 0B94 TAMIL LETTER AU appears as if it is constructed from the graphic symbols for 0B93 TAMIL LETTER OO and 0BD7 TAMIL AU LENGTH MARK

EXAMPLE 2 Row 0D Malayalam

The graphic symbol for 0D4A MALAYALAM VOWEL SIGN O appears as if it is constructed from the graphic symbols for 0D46 MALAYALAM VOWEL SIGN E and 0D3E MALAYALAM VOWEL SIGN AA

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, shall be regarded as equivalent to a sequence of two or more other coded characters taken from the same table.

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.

23 Source references for CJK Ideographs

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 – 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 in 23.1 and 23.3. Such a source standard originates from the following categories:

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

23.1 Source references for CJK Unified Ideographs

The procedures that were used to derive the unified ideographs from the source character set standards, and the rules for their arrangement in the code charts in 30, are described in Annex S.

NOTE 1 – The source separation rule described by the clause S.1.6 of that annex only apply to CJK Unified Ideographs within the BMP.

The following list identifies all sources referenced by the CJK Unified Ideographs in both the BMP and the SIP. The current full set of CJK Unified Ideographs is represented by the collection 387 CJK UNIFIED IDEOGRAPHS-2009 (See A.1).

The Hanzi G sources are

G0	GB2312-80
G1	GB12345-90 with 58 Hong Kong and 92 Korean “Idu” characters
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
G_4K	Siku Quanshu（四庫全書）
G_BK	Chinese Encyclopedia（中國大百科全書）
G_CH	Ci Hai（辞海）
G_CY	Ci Yuan（辭源）
G_CYY	Chinese Academy of Surveying and Mapping Ideographs (中国测绘科学院用字）
G_FZ	Founder Press System（方正排版系统）
G_GH	Gudai Hanyu Cidian（古代汉语词典）
G_GJZ	Commercial Press Ideographs（商务印书馆用字）
G_HC	Hanyu Dacidian（漢語大詞典）
G_HZ	Hanyu Dazidian ideographs（漢語大字典）
G_KX	Kangxi Dictionary ideographs（康熙字典）9th edition (1958) including the addendum（康熙字典）補遺
G_XC	Xiandai Hanyu Cidian (现代汉语词典）
G_ZFY	Hanyu Fangyan Dacidian（汉语方言大辞典）
G_ZJW	Yinzhou Jinwen Jicheng Yinde（殷周金文集成引得）

The Hanzi H source is

H	Hong Kong Supplementary Character Set – 2004

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
TC	TCA-CNS 11643-1992 12th plane
TD	TCA-CNS 11643-1992 13th plane
TE	TCA-CNS 11643-1992 14th plane
TF	TCA-CNS 11643-1992 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
JK	Japanese KOKUJI Collection
J_ARIB	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
K1	KS C 5657-1991
K2	PKS C 5700-1 1994
K3	PKS C 5700-2 1994
K4	PKS 5700-3:1998
K5	Korean IRG Hanja Character Set 5th Edition: 2001

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
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, June 2008

NOTE 2 – Even if source references get updated, the source reference information is not updated. The updated source references may only identify characters not previously covered by the older version.

The content linked to is a plain text file, using ISO/IEC 646-IRV characters with LINE FEED as end of line mark, that specifies, after a 13-lines header, as many lines as CJK Unified Ideographs in the sum of the two planes; each containing the following information organized in fields delimited by ';' (empty fields use no character):

- 1st field: BMP or SIP code point `(0hhhh), (2hhhh)`
- 2nd field: Radical Stroke index `(d{1,3}'.d{1,2})`, (Radical is one to three digits, optionally followed by an apostrophe for alternate radical, followed by a full stop, and ending by one or two digits for the stroke count).
- 3rd field: Hanzi G sources `(G0-hhhh), (G1-hhhh), (G3-hhhh), (G5-hhhh), (G7-hhhh), (GS-hhhh), (G8-hhhh), (G9-hhhh), (GE-hhhh), (G_KXdddddd), (G_HZ), (G_HZddddd), (G_CY), (G_CH), (G_CHddddd), (G_HC), (G_HCddddd), (G_BK), (G_BKddddd), (G_FZ), (G_FZddddd), (G_4K), (G_GHddddd), (G_GJZddddd), (G_XCddddd), (G_CYYddddd), (G_ZFYddddd),` or `(G_ZJWddddd)`
- 4th field: Hanzi T sources `T1-hhhh), (T2-hhhh), (T3-hhhh), (T4-hhhh), (T5-hhhh), (T6-hhhh), (T7-hhhh), (TC-hhhh), (TD-hhhh), (TE-hhhh),` or `(TF-hhhh)`
- 5h field: Kanji J sources `(J0-hhhh), (J1-hhhh), (J3-hhhh), (J3A-hhhh), (J4-hhhh), (JA-hhhh), (JK-ddddd),` or `(J_ARIB-hhhh)`
- 6th field: Hanja K sources `(K0-hhhh), K1-hhhh), (K2-hhhh), (K3-hhhh),(K4-hhhh),` or `(K5-hhhh)`
- 7th field: ChuNom V sources `(V0-hhhh), V1-hhhh), (V2-hhhh), (V3-hhhh),` or `(V4-hhhh)`
- 8th field: Hanzi H source `(H-hhhh)`
- 9th field: Hanja KP sources `(KP0-hhhh)` or `(KP1-hhhh)`
- 10th field: Unicode U source `(UTCddddd)`
- 11th field: Hanzi M source `(MACddddd)`

The format definition uses `'d'` as a decimal unit and `'h'` as a hexadecimal unit. Uppercase characters, digits and all other symbols between parentheses appear as shown.

NOTE 3 – 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 4 – The original source references in the Hanja K4 source (PKS 5700-3:1998) 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 the other hexadecimal values, they do not decompose in section, position values.

Click on this highlighted text to access the reference file.

NOTE 5 – The content is also available as a separate viewable file in the same file directory as this document. The file is named: "CJKU_SR.txt".

23.2 Source reference presentation for CJK Unified Ideographs

Unlike 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.

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 source references numbers are represented using the format specified in 23.1 with two minor variations:

- G_KX sources use the format `'KXdddd.dd'` instead of `'G_KXdddddd'`.

- J_ARIB sources use the format ‘JARIB-hhhh’ instead of ‘J_ARIB-hhhh’.

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, V, KP, H, U, and M.

23.2.1 Source reference presentation for CJK UNIFIED IDEOGRAPHS block

The following figure shows the presentation for the CJK UNIFIED IDEOGRAPH block. Graphic representations for the Hanzi G and T sources, the Kanji J source, the Hanja K source, the Hanja KP source, and the ChuNom V source are shown in that order when present. If sources from another of the two Hanzi sources (H or M sources) also exist, they are shown in a new row in the same Hanzi column. U sources are shown in the same column as J sources but in a new row.

The following figure shows an example for characters 4E00-4E09 and 4E12-4E1A.

HEX	C	J	K	V
4E00	一 一	一	一	一 一 一
— 1.0	G0-523B T1-4421	J0-306C	K0-6C69	KP0-FCD6 V1-4A21
4E01	丁 丁	丁	丁	丁 丁 丁
— 1.1	G0-3621 T1-4421	J0-437A	K0-6F4B	KP0-E8B9 V1-4A22
4E02	亏 亏	亏		
— 1.1	G5-3021 T4-2126	J1-3021		
4E03	七 七	七	七	七 七 七
— 1.1	G0-465F T1-4424	J0-3C37	K0-7652	KP0-EFA6 V1-4A23
4E04	丄 丄	丄		
— 1.1	G0-523B T1-4421	J0-306C		
	丄			
	H-9E93			
4E05	丅 丅	丅		
— 1.1	GE-2122 T3-2125	J1-3023		
4E06	丆		丆	
— 1.1	G1-7D3D		K2-2121	
4E07	万 万	万	万	万 万 万
— 1.2	G0-4D72 T2-2126	J0-4B7C	K0-5832	KP0-DAB9 V1-4A24
4E08	丈 丈	丈	丈	丈 丈 丈
— 1.2	G0-5549 T1-4437	J0-3E66	K0-6D5B	KP0-E6DD V1-4A25
4E09	三 三	三	三	三 三 三
— 1.2	G0-487D T1-4435	J0-3B30	K0-5F32	KP0-E1B5 V1-4A26

HEX	C	J	K	V
4E12	刃 刃	刃		刃
— 1.3	GE-2123 T4-2139	J1-3025		KP1-340E
4E13	专			
— 1.3	G0-5728			
4E14	且 且	且	且	且 且 且
— 1.4	G0-4752 T1-4562	J0-336E	K0-7326	KP0-ECA8 V1-4A2D
4E15	丕 丕	丕	丕	丕 丕 丕
— 1.4	G0-5287 T1-4561	J0-5023	K0-5D60	KP0-DFC9 V1-4A2E
4E16	世 世	世	世	世 世 世
— 1.4	G0-4A40 T1-4560	J0-4024	K0-6126	KP0-E5F9 V1-4A2F
4E17	丗 丗	丗		丗
— 1.3	GE-2124 T4-2155	J0-5242		KP1-3413
4E17	丗 丗	丗		丗
— 1.3	GE-2124 T4-2155	J0-5242		KP1-3413
4E18	丘 丘	丘	丘	丘 丘 丘
— 1.4	G0-4770 T1-4563	J0-3556	K0-4E78	KP0-D0DF V1-4A30
4E19	丙 丙	丙	丙	丙 丙 丙
— 1.4	G0-317B T1-455F	J0-4A3A	K0-5C30	KP0-DDF9 V1-4A31
4E1A	业			
— 1.4	G0-5235			
	业			
	H-9EB2			

Figure 2 – Source reference presentation for CJK UNIFIED IDEOGRAPHS

23.2.2 Source reference presentation for CJK UNIFIED IDEOGRAPHS EXTENSION A

The following figure shows the presentation for the CJK UNIFIED IDEOGRAPH EXTENSION A block. Up to four sources per characters are represented in a single row. If more than four sources exist, an additional row is used.

The following figure shows an example for characters 41CB-41CC, 41DC, and 41EE.

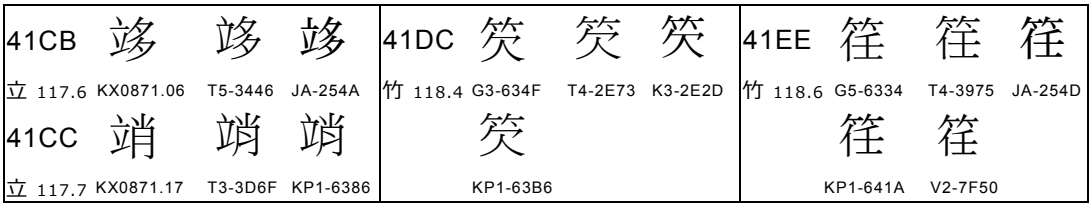


Figure 3 – Source reference presentation for CJK UNIFIED IDEOGRAPHS EXTENSION A

23.2.3 Source reference presentation for CJK UNIFIED IDEOGRAPHS EXTENSION B and C

The following figure shows the presentation for the CJK UNIFIED IDEOGRAPH EXTENSION B and C 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 following figure shows an example for characters 2000F-20010, 200021, 20032-20033, and 20043-20044.

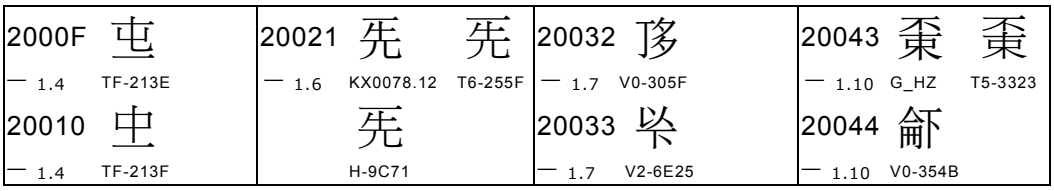


Figure 4 – Source reference presentation for CJK UNIFIED IDEOGRAPHS EXTENSION B and C

23.3 Source references for CJK Compatibility Ideographs

The following list identifies all sources referenced by the CJK Compatibility Ideographs in both the BMP and the SIP. The current full set of CJK Compatibility Ideographs is represented by the collection 386 CJK COMPATIBILITY IDEOGRAPHS-2008 (See A.1).

The Hanzi H source is

H Hong Kong Supplementary Character Set - 2004

Hanzi T sources are

T3 TCA-CNS 11643-1992 3rd plane
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
TF TCA-CNS 11643-1992 15th plane

Kanji J sources are

J3 JIS X 0213:2000 level-3
J4 JIS X 0213:2000 level-4
J_ARIB Association of Radio Industries and Businesses (ARIB) ARIB STD-B24 Version 5.1, March 14 2007

The Hanja K source is

K0 KS C 5601-1987

The Hanja KP source is

KP1 KPS 10721-2000

The Unicode U source is

UTC The Unicode Technical Report #45, U-source Ideographs, June 2008

The content linked to is a plain text file, using ISO/IEC 646-IRV characters with LINE FEED as end of line mark, that specifies, after a 11-lines header, as many lines as CJK Compatibility Ideographs; each containing the following information organized in fields delimited by ';' (empty fields use no character):

- 1st field: BMP or SIP code point `(0hhhh )` or `(2hhhh)`
- 2nd field: Code point of corresponding CJK Unified Ideograph `(0hhhh)` or `(2hhhh)`
- 3rd field: Hanzi T sources `(T3-hhhh), (T4-hhhh), (T5-hhhh), (T6-hhhh), (T7-hhhh),` or `(TF-hhhh)`
- 4th field: Hanzi H source `(H-hhhh)`
- 5th field: Kanji J sources `(J3-hhhh), J4-hhhh)`
- 6th field: Hanja K source `(K0-hhhh)`
- 7th field: Unicode U source `(UTCddddd)`
- 8th field: Hanja KP source `(KP1-hhhh)`

The format definition uses 'h' as a hexadecimal unit. Uppercase characters, digits and all other symbols between parentheses appear as shown.

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.

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: "CJKC_SR.txt".

24 Character names and annotations

24.1 Entity names

This standard specifies names for the following entity types

- characters
- named UCS sequences identifiers (see 25)
- blocks (see 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 this clause. 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 names 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. 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.

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.

24.4 Name uniqueness

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

24.4.1 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.

24.4.2 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.

24.4.3 Character names and named UCS sequence identifiers

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

24.4.4 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.

NOTE 1 – 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 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 2 – 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.

24.5 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”.

24.6 Character names for Tangut characters

For Tangut characters the names are algorithmically constructed by appending their coded representation in hexadecimal notation using their five hexadecimal digit value to “TANGUT CHARACTER-”.

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.

1) 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.

2) 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.

3) Calculate the character index C from the formula
$C = 4096 \times (d_1 - 10) + 256 \times (d_2 - 12) + 16 \times d_3 + d_4$

4) Calculate the syllable component indices I, P, F from the following formulae
$I = C / 588$ (Note: $0 \le I \le 18$)
$P = (C \% 588) / 28$ (Note: $0 \le P \le 20$)
$F = C \% 28$ (Note: $0 \le F \le 27$)
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).

5) 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.

6) 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 5.

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 short additional information is defined and available in annex Annex R along with each character name. This additional information 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.

7) Carry out steps 1 to 4 as described above.

8) Obtain the Latin character strings that correspond to the three indices I, P, F from columns 5, 6, and 7 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.

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 additional information is phwips.

NOTE – The annex R provides the names of Hangul syllables through a linked file.

Table 5: Elements of Hangul syllable names and additional information

Index number	Syllable name elements			Additional information		
	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

25 Named UCS Sequence Identifiers

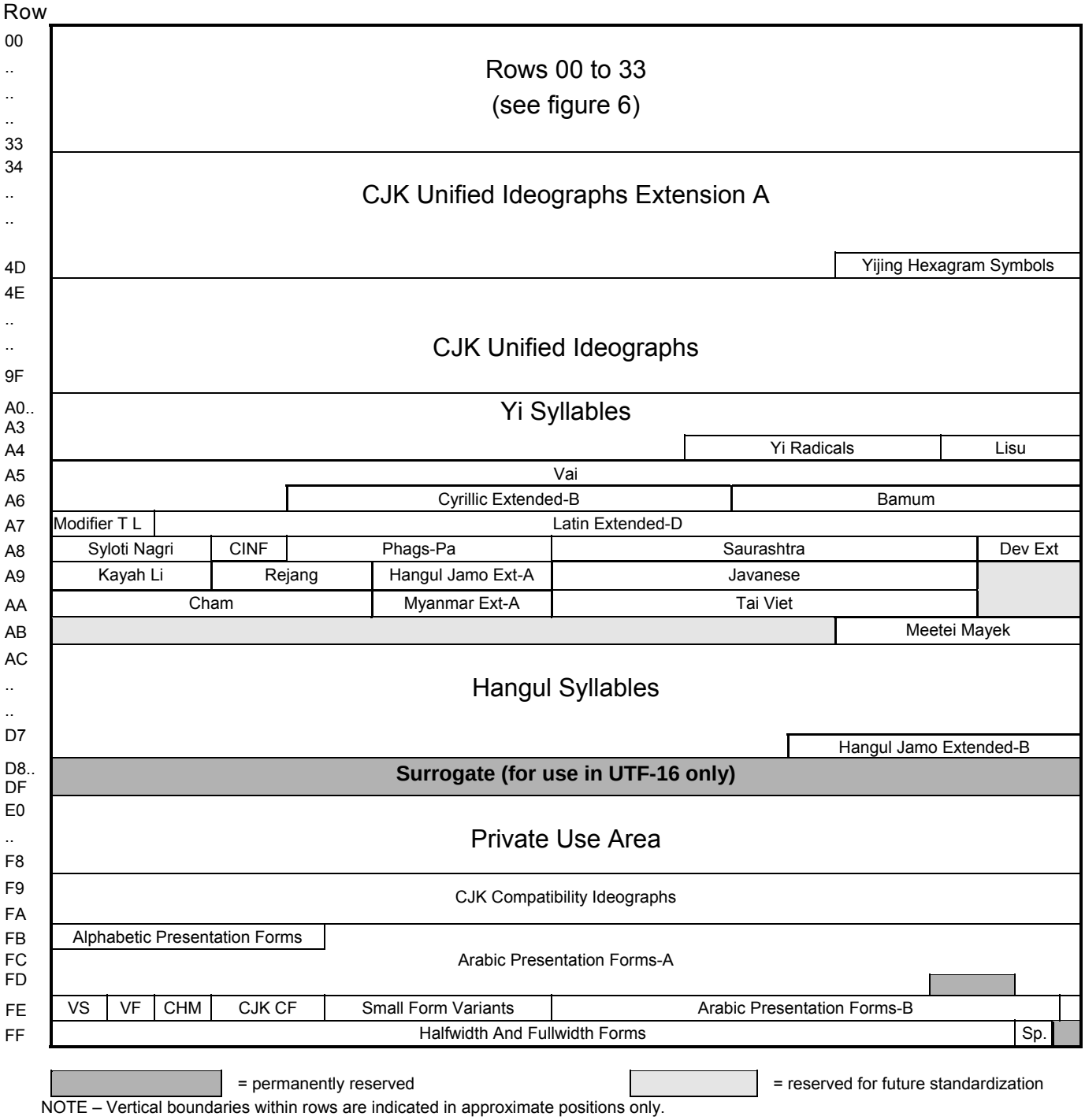
A Named UCS Sequence Identifier (NUSI) is a USI associated to a name following the same construction rules as for character names. These rules are given in 24.

NOTE – 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 NFC (see 21). Each named UCS sequence has a unique code representation. All the allowed named UCS sequence identifiers for use with ISO/IEC 10646 are specified in the Unicode Standard UAX#34 (see 3); all other such named sequences are undefined.

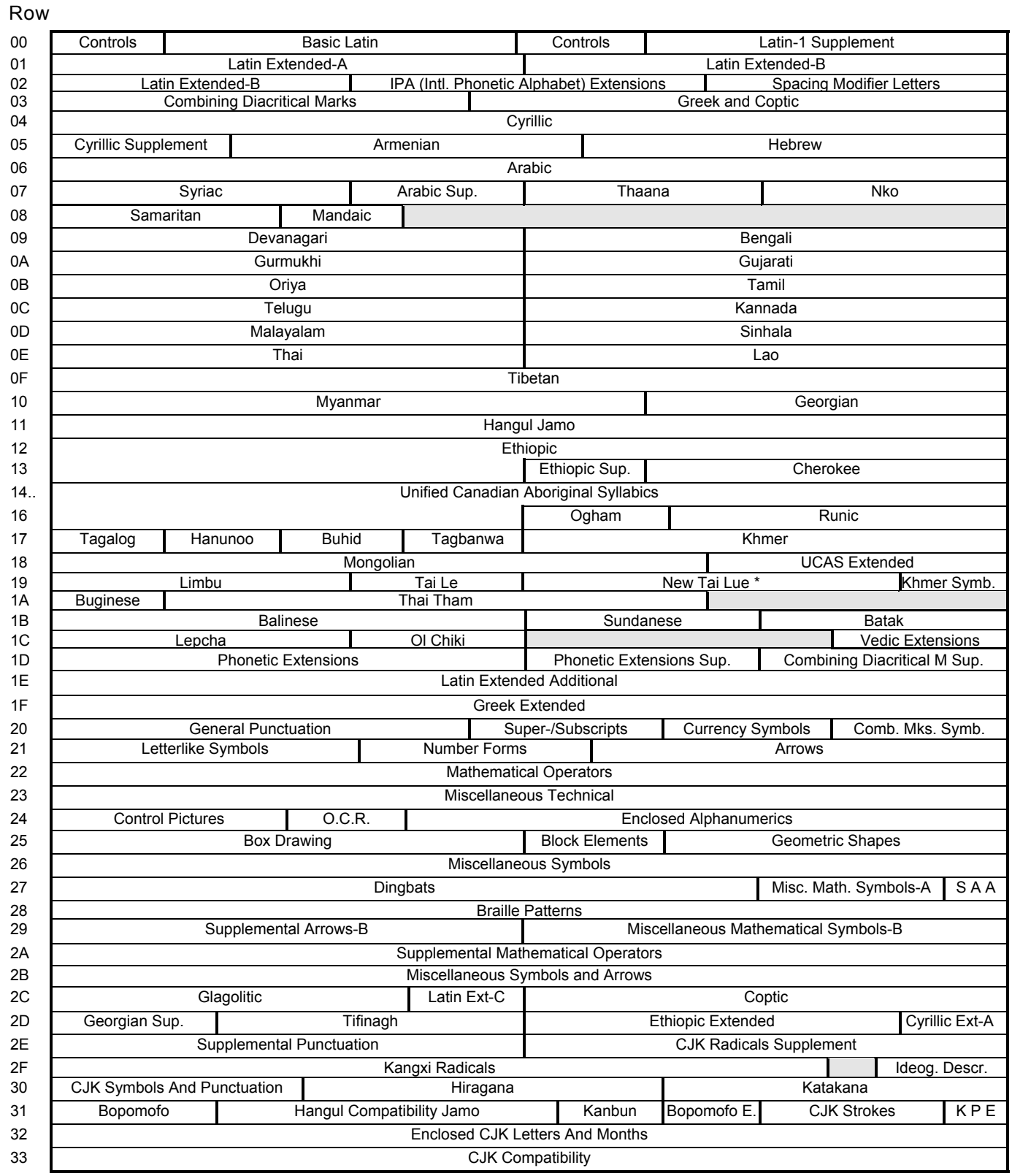
26 Structure of the Basic Multilingual Plane

An overview of the Basic Multilingual Plane is shown in figure 3 and a more detailed overview of Rows 00 to 33 is shown in figure 4. The Basic Multilingual Plane includes characters in general use in alphabetic, syllabic, and ideographic scripts together with various symbols and digits.



NOTE – Vertical boundaries within rows are indicated in approximate positions only.

Figure 5 - Overview of the Basic Multilingual Plane



* NOTE 1 – New Tai Lue is also known as Xishuang Banna Dai

NOTE 2 – Vertical boundaries within rows are indicated in approximate positions only.

Figure 6 - Overview of Rows 00 to 33 of the Basic Multilingual Plane

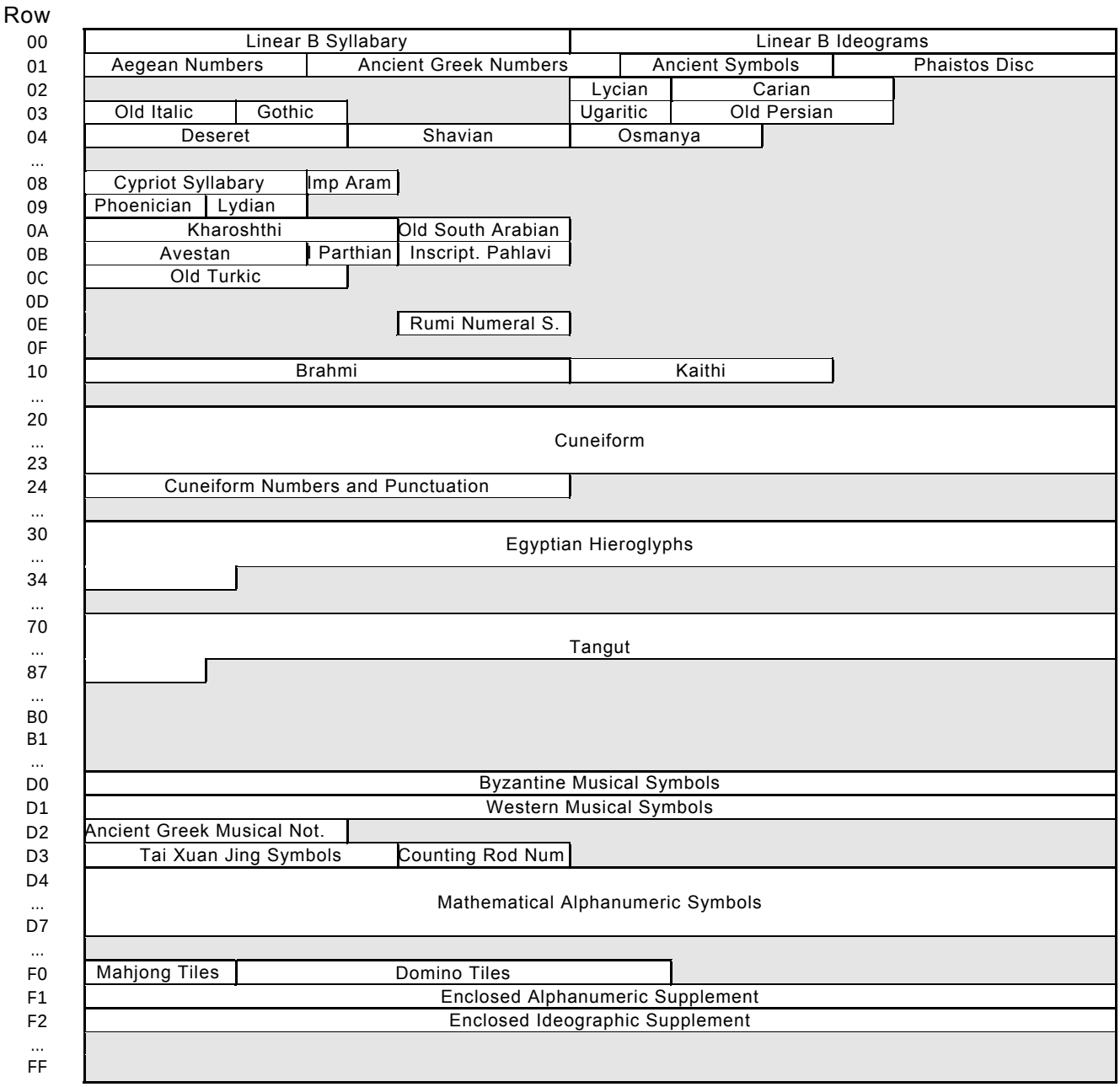
27 Structure of the Supplementary Multilingual Plane for scripts and symbols (SMP)

Because another supplementary plane is reserved for additional CJK Ideographs, the SMP (plane 1) 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 1 – 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 5.



= reserved for future standardization

NOTE 2 – Vertical boundaries within rows are indicated in approximate positions only.

NOTE 3 – The Old Italic block represents a unified script that covers the Etruscan, Oscan, Umbrian, Faliscan, North Picene, and South Picene alphabets. Some of these alphabets can be written with characters oriented in either left-to-right or right-to-left direction. The glyphs in the code chart are shown with left to right orientation.

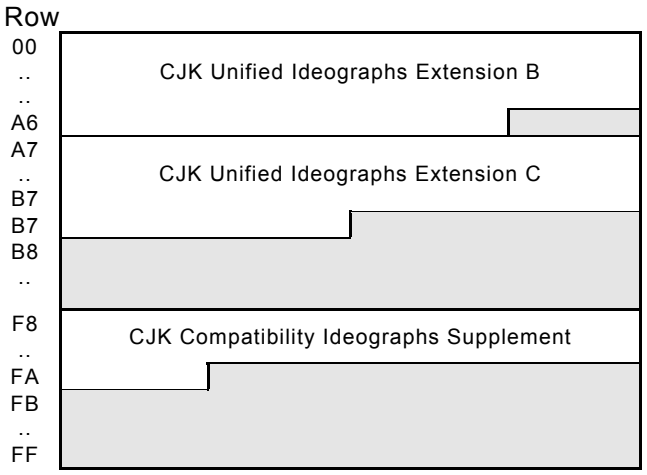
Figure 7 – Overview of the Supplementary Multilingual Plane for scripts and symbols

28 Structure of the Supplementary Ideographic Plane (SIP)

The SIP (plane 2) 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 compatibility CJK ideographs. These ideographs are compatibility characters as specified in 18.

The following figure 6 shows an overview of the Supplementary Ideographic Plane.



NOTE – Vertical boundaries within rows are indicated in approximate positions only.

Figure 8 – Overview of the Supplementary Ideographic Plane

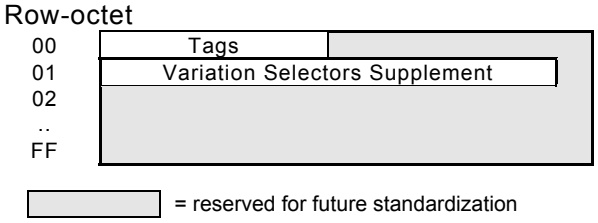
29 Structure of the Supplementary Special-purpose Plane (SSP)

The SSP (plane 0E) is used for special purpose use graphic characters. Code points from E0000 to E0FFF are reserved for Format Characters (see 16).

NOTE 1 – Some of these characters do not have a visual representation and do not have printable graphic symbols. The Tag Characters are example of such characters.

An overview of the Supplementary Special-purpose Plane is shown in figure 7.

NOTE 2 – Unassigned code points in this range should be ignored in normal processing and display.



NOTE 3 – Vertical boundaries within rows are indicated in approximate positions only.

Figure 9 – Overview of the Supplementary Special-purpose Plane

30 Code charts and lists of character names

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 the CJK UNIFIED IDEOGRAPHS blocks, the HANGUL SYLLABLES block, and the TANGUT block.

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

30.1 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.

30.2 Character names list

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

- Character code position,
- Associated character name,
- Character alias (one preceded by '※') which is a unique and stable alternate name for a character.

All other information is informative and may contain:

- Graphic symbol associated with the character
- Subheads grouping various subsets 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 alias 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.
- Decompositions, preceded by '≡', or '≈', describing various mapping between characters.

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 lj LATIN SMALL LETTER IJ

→ 0459 љ cyrillic small letter lje

≈ 006C l 006A j

...

FE18 ﹘ PRESENTATION FORM FOR VERTICAL RIGHT WHITE LENTICULAR BRAKCET

※ PRESENTATION FORM FOR VERTICAL RIGHT WHITE LENTICULAR BRACKET

• misspelling of “BRACKET” in character name is a known defect

≈ <vertical> 3017 〗

30.3 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 CJK Compatibility (0000-33FF)
- CJK Unified Ideographs Extension A (3400-4DBF)
- Yijing Hexagram Symbols (4DC0-4DFF)
- CJK Unified Ideographs Part 1 of 3 (4E00-6912)
- CJK Unified Ideographs Part 2 of 3 (6913-842D)
- CJK Unified Ideographs Part 3 of 3 (842E-9FFF)
- Yi Syllables to Specials (A000-FFFF)
- Linear B Syllabary to Enclosed Ideographic Supplement (10000-1FFFF)
- CJK Unified Ideographs Extension B Part 1 of 3 (20000-23AF1)
- CJK Unified Ideographs Extension B Part 2 of 3 (23AF2-272AB)
- CJK Unified Ideographs Extension B Part 3 of 3 (272AC-2A6FF)
- CJK Unified Ideographs Extension C (2A700-2B73F)
- CJK Compatibility Ideographs (2F800-DFFFF)
- Tag to Variation Selectors Supplement (E0000-10FFFF)

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 *
2	LATIN-1 SUPPLEMENT	00A0-00FF *
3	LATIN EXTENDED-A	0100-017F *
4	LATIN EXTENDED-B	0180-024F *
5	IPA EXTENSIONS	0250-02AF *
6	SPACING MODIFIER LETTERS	02B0-02FF *
7	COMBINING DIACRITICAL MARKS	0300-036F *
8	BASIC GREEK	0370-03CF
9	GREEK SYMBOLS AND COPTIC	03D0-03FF
10	CYRILLIC	0400-04FF *
11	ARMENIAN	0530-058F
12	BASIC HEBREW	05D0-05EA *
13	HEBREW EXTENDED	0590-05CF 05EB-05FF
14	BASIC ARABIC	0600-065F
15	ARABIC EXTENDED	0660-06FF *
16	DEVANAGARI	0900-097F 200C, 200D
17	BENGALI	0980-09FF 200C, 200D
18	GURMUKHI	0A00-0A7F 200C, 200D
19	GUJARATI	0A80-0AFF 200C, 200D
20	ORIYA	0B00-0B7F 200C, 200D
21	TAMIL	0B80-0BFF 200C, 200D
22	TELUGU	0C00-0C7F 200C, 200D
23	KANNADA	0C80-0CFF 200C, 200D
24	MALAYALAM	0D00-0D7F 200C, 200D
25	THAI	0E00-0E7F
26	LAO	0E80-0EFF
27	BASIC GEORGIAN	10D0-10FF
28	GEORGIAN EXTENDED	10A0-10CF
29	HANGUL JAMO	1100-11FF *
30	LATIN EXTENDED ADDITIONAL	1E00-1EFF *
31	GREEK EXTENDED	1F00-1FFF
32	GENERAL PUNCTUATION	2000-206F
33	SUPERSCRIPTS AND SUBSCRIPTS	2070-209F
34	CURRENCY SYMBOLS	20A0-20CF
35	COMBINING DIACRITICAL MARKS FOR SYMBOLS	20D0-20FF
36	LETTERLIKE SYMBOLS	2100-214F *
37	NUMBER FORMS	2150-218F
38	ARROWS	2190-21FF *
39	MATHEMATICAL OPERATORS	2200-22FF *
40	MISCELLANEOUS TECHNICAL	2300-23FF
41	CONTROL PICTURES	2400-243F
42	OPTICAL CHARACTER RECOGNITION	2440-245F
43	ENCLOSED ALPHANUMERICS	2460-24FF *
44	BOX DRAWING	2500-257F *
45	BLOCK ELEMENTS	2580-259F *
46	GEOMETRIC SHAPES	25A0-25FF *
47	MISCELLANEOUS SYMBOLS	2600-26FF *
48	DINGBATS	2700-27BF
49	CJK SYMBOLS AND PUNCTUATION	3000-303F *
50	HIRAGANA	3040-309F
51	KATAKANA	30A0-30FF *
52	BOPOMOFO	3100-312F 31A0-31BF
53	HANGUL COMPATIBILITY JAMO	3130-318F
54	CJK MISCELLANEOUS	3190-319F
55	ENCLOSED CJK LETTERS AND MONTHS	3200-32FF
56	CJK COMPATIBILITY	3300-33FF *
57, 58, 59 (These collection numbers shall not be used, see Note 2.)		
60	CJK UNIFIED IDEOGRAPHS	4E00-9FFF
61	PRIVATE USE AREA	E000-F8FF
62	CJK COMPATIBILITY IDEOGRAPHS	F900-FAFF
63	(Collection specified as union of other collections)	
64	ARABIC PRESENTATION FORMS-A	FB50-FDCF FDF0-FDFF
65	COMBINING HALF MARKS	FE20-FE2F
66	CJK COMPATIBILITY FORMS	FE30-FE4F *
67	SMALL FORM VARIANTS	FE50-FE6F
68	ARABIC PRESENTATION FORMS-B	FE70-FEFE
69	HALFWIDTH AND FULLWIDTH FORMS	FF00-FFEF
70	SPECIALS	FFF0-FFFD
71	HANGUL SYLLABLES	AC00-D7A3 *
72	BASIC TIBETAN	0F00-0FBF
73	ETHIOPIC	1200-137F

74	UNIFIED CANADIAN ABORIGINAL SYLLABICS	1400-167F *
75	CHEROKEE	13A0-13FF
76	YI SYLLABLES	A000-A48F
77	YI RADICALS	A490-A4CF
78	KANGXI RADICALS	2F00-2FDF
79	CJK RADICALS SUPPLEMENT	2E80-2EFF
80	BRAILLE PATTERNS	2800-28FF
81	CJK UNIFIED IDEOGRAPHS EXTENSION A	3400-4DBF FA1F, FA23
82	OGHAM	1680-169F
83	RUNIC	16A0-16FF
84	SINHALA	0D80-0DFF
85	SYRIAC	0700-074F
86	THAANA	0780-07BF
87	BASIC MYANMAR	1000-104F * 200C, 200D
88	KHMER	1780-17FF 200C, 200D
89	MONGOLIAN	1800-18AF
90	EXTENDED MYANMAR	1050-109F *
91	TIBETAN	0F00-0FFF
92	CYRILLIC SUPPLEMENT	0500-052F
93	TAGALOG	1700-171F
94	HANUNOO	1720-173F
95	BUHID	1740-175F
96	TAGBANWA	1760-177F
97	MISCELLANEOUS MATHEMATICAL SYMBOLS-A	27C0-27EF
98	SUPPLEMENTAL ARROWS-A	27F0-27FF *
99	SUPPLEMENTAL ARROWS-B	2900-297F *
100	MISCELLANEOUS MATHEMATICAL SYMBOLS-B	2980-29FF *
101	SUPPLEMENTAL MATHEMATICAL OPERATORS	2A00-2AFF *
102	KATAKANA PHONETIC EXTENSIONS	31F0-31FF *
103	VARIATION SELECTORS	FE00-FE0F *
104	LTR ALPHABETIC PRESENTATION FORMS	FB00-FB1C
105	RTL ALPHABETIC PRESENTATION FORMS	FB1D–FB4F
106	LIMBU	1900-194F
107	TAI LE	1950-197F
108	KHMER SYMBOLS	19E0-19FF *
109	PHONETIC EXTENSIONS	1D00-1D7F *
110	MISCELLANEOUS SYMBOLS AND ARROWS	2B00-2BFF
111	YIJING HEXAGRAM SYMBOLS	4DC0-4DFF *
112	ARABIC SUPPLEMENT	0750-077F *
113	ETHIOPIC SUPPLEMENT	1380-139F
114	NEW TAI LUE	1980-19DF
115	BUGINESE	1A00-1A1F
116	PHONETIC EXTENSIONS SUPPLEMENT	1D80-1DBF *
117	COMBINING DIACRITICAL MARKS SUPPLEMENT	1DC0-1DFF
118	GLAGOLITIC	2C00-2C5F
119	COPTIC	03E2-03EF 2C80-2CFF
120	GEORGIAN SUPPLEMENT	2D00-2D2F
121	TIFINAGH	2D30-2D7F
122	ETHIOPIC EXTENDED	2D80-2DDF
123	SUPPLEMENTAL PUNCTUATION	2E00-2E7F
124	CJK STROKES	31C0-31EF
125	MODIFIER TONE LETTERS	A700-A71F *
126	SYLOTI NAGRI	A800-A82F
127	VERTICAL FORMS	FE10-FE1F
128	NKO	07C0-07FF
129	BALINESE	1B00-1B7F
130	LATIN EXTENDED-C	2C60-2C7F *
131	LATIN EXTENDED-D	A720-A7FF
132	PHAGS-PA	A840-A87F
133	SUNDANESE	1B80-1BBF
134	LEPCHA	1C00-1C4F
135	OL CHIKI	1C50-1C7F *
136	VAI	A500-A63F
137	SAURASHTRA	A880-A8DF
138	KAYAH LI	A900-A92F *
139	REJANG	A930-A95F
140	CYRILLIC EXTENDED-A	2DE0-2DFF *
141	CYRILLIC EXTENDED-B	A640-A69F
142	CHAM	AA00-AA5F
143	TAI THAM	1A20-1AAF
144	HANGUL JAMO EXTENDED-A	A960-A97F
145	TAI VIET	AA80-AADF
146	HANGUL JAMO EXTENDED-B	D7B0-D7FF
147	SAMARITAN	0800-083F
148	UNIFIED CANADIAN ABORIGINAL SYLLABICS EXTENDED	18B0-18FF
149	VEDIC EXTENSIONS	1CD0-1CFF
150	LISU	A4D0-A4FF *
151	BAMUM	A6A0-A6FF
152	COMMON INDIC NUMBER FORMS	A830-A83F
153	DEVANAGARI EXTENDED	A8E0-A8FF
154	JAVANESE	A980-A9DF
155	MYANMAR EXTENDED-A	AA60-AA7F
156	MEETEI MAYEK	ABC0-ABFF
157	MANDAIC	0840-085F
158	BATAK	1BC0-1BFF
1001	OLD ITALIC	10300-1032F
1002	GOTHIC	10330-1034F
1003	DESERET	10400-1044F *
1004	BYZANTINE MUSICAL SYMBOLS	1D000-1D0FF
1005	MUSICAL SYMBOLS	1D100-1D1FF

1006	MATHEMATICAL ALPHANUMERIC SYMBOLS	1D400-1D7FF
1007	LINEAR B SYLLABARY	10000-1007F
1008	LINEAR B IDEOGRAMS	10080-100FF
1009	AEGEAN NUMBERS	10100-1013F
1010	UGARITIC	10380-1039F
1011	SHAVIAN	10450-1047F *
1012	OSMANYA	10480-104AF
1013	CYPRIOT SYLLABARY	10800-1083F
1014	TAI XUAN JING SYMBOLS	1D300-1D35F
1015	ANCIENT GREEK NUMBERS	10140-1018F
1016	OLD PERSIAN	103A0-103DF
1017	KHAROSHTHI	10A00-10A5F
1018	ANCIENT GREEK MUSICAL NOTATION	1D200-1D24F
1019	PHOENICIAN	10900-1091F
1020	CUNEIFORM	12000-123FF
1021	CUNEIFORM NUMBERS AND PUNCTUATION	12400-1247F
1022	COUNTING ROD NUMERALS	1D360-1D37F
1023	PHAISTOS DISC	101D0-101FF
1024	LYCIAN	10280-1029F
1025	CARIAN	102A0-102DF
1026	LYDIAN	10920-1093F
1027	ANCIENT SYMBOLS	10190-101CF
1028	MAHJONG TILES	1F000-1F02F
1029	DOMINO TILES	1F030-1F09F
1030	AVESTAN	10B00-10B3F
1031	EGYPTIAN HIEROGLYPHS	13000-1342F
1032	IMPERIAL ARAMAIC	10840-1085F
1033	OLD SOUTH ARABIAN	10A60-10A7F
1034	INSCRIPTIONAL PARTHIAN	10B40-10B5F
1035	INSCRIPTIONAL PAHLAVI	10B60-10B7F
1036	OLD TURKIC	10C00-10C4F
1037	RUMI NUMERAL SYMBOLS	10E60-10E7F
1038	KAITHI	11080-110CF
1039	ENCLOSED ALPHANUMERIC SUPPLEMENT	1F100-1F1FF
1040	ENCLOSED IDEOGRAPHIC SUPPLEMENT	1F200-1F2FF
1041	BRAHMI	11000-1107F
1042	TANGUT	17000-1871F
2001	CJK UNIFIED IDEOGRAPHS EXTENSION B	20000-2A6DF
2002	CJK COMPATIBILITY IDEOGRAPHS SUPPLEMENT	2F800-2FA1F
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	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.3.2) *	
301	BMP-AMD.7	see A.3.1 *
302	BMP SECOND EDITION	see A.3.3 *
303	UNICODE 3.1	see A.6.1 *
304	UNICODE 3.2	see A.6.2 *
305	UNICODE 4.0	see A.6.3 *
306	UNICODE 4.1	see A.6.4 *
307	UNICODE 5.0	see A.1.1A.6.5 *
308	UNICODE 5.1	see A.1.1A.6.6 *
309	UNICODE 5.2	see A.1.1A.6.7 *
310	UNICODE 6.0	see *

340	COMBINED FIRST EDITION	see A.1.1A.3.4 *
10646	UNICODE	0000-FDCF FDF0-FFFD 10000-1FFFD 20000-2FFFD 30000-3FFFD 40000-4FFFD 50000-5FFFD 60000-6FFFD 70000-7FFFD 80000-8FFFD 90000-9FFFD A0000-AFFFD B0000-BFFFD C0000-CFFFD D0000-DFFFD E0000-EFFFD F0000-FFFFD 100000-10FFFD

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

The following specify other collections, including extended collections.

270	COMBINING CHARACTERS	BMP characters specified in clause 4.14
271	(This collection number shall not be used, see Note 2) *	
281	MES-1	see A.5.1 *
282	MES-2	see A.5.2 *
283	MODERN EUROPEAN SCRIPTS	see A.5.3 *
284	CONTEMPORARY LITHUANIAN LETTERS	see A.1.1A.5.4 *
285	BASIC JAPANESE	see A.5.5 *
286	JAPANESE NON IDEOGRAPHICS EXTENSION	see A.5.6 *
287	COMMON JAPANESE	see A.1.1A.5.7 *
300	BMP	0000-D7FF E000-FFFD
400	(This collection number shall not be used, see Note 3.)	
401	PRIVATE USE PLANES-0F-10	G=00, P=0F-10
500	(This collection number shall not be used, see Note 3.)	
1000	SMP	10000-1FFFD
1900	SMP COMBINING CHARACTERS	SMP characters specified in clause 4.14
2000	SIP	20000-2FFFD
3000	SSP	E0000-EFFFD

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 ISO/IEC 10646 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
Alphabetic	63
Alphanumeric	43
Ancient Greek	1015 1018
Arabic	14 15 64 68 112
Armenian	11
Arrows	38 98 99 110
Avestan	1030
Balinese	129
Bamum	151
Batak	158
Bengali	17
Bidirectional	202 203
Block elements	45
BMP	300 301 302 (299)
Box drawing	44
Bopomofo	52
Brahmi	1041
Braille patterns	80
Buginese	115
Buhid	95
Byzantine musical symbols	1004
Canadian Aboriginal	74 148
Carian	1025
Cham	142
Cherokee	75
CJK	49 54 55 56 60 62 66 78 81 124 2001 2002
Combining	7 35 65 117 270 271
Compatibility	53 56 62 66
Control pictures	41
Coptic	9 119
Counting Rod numerals	1022
Cuneiform	1020 1021
Currency	34
Cypriot syllabary	1013
Cyrillic	10 92 140 141
Deseret	1003
Devanagari	16 153
Diacritical marks	7 35 117
Dingbats	48
Enclosed	43 55
Egyptian Hieroglyphs	1031
Ethiopic	73 113 122
Format	201 202 203 250 251
Fullwidth	69
Game Tiles	1028, 1029
Geometric shapes	46
Georgian	27 28 120
Glagolitic	118
Gothic	1002
Greek	8 9 31
Gujarati	19
Gurmukhi	18
Half (marks, width)	65 69
Hangul	29 53 71 144 146 204
Hanunoo	94
Hebrew	12 13
Hiragana	50
Ideographs	60 62 81 207 380 381
Imperial Aramaic	1032
Inscriptional Pahlavi	1035
Inscriptional Parthian	1034
IPA extensions	5
Jamo	29 53 144 146
Javanese	154
Kaithi	1038
Kangxi	78
Kannada	23
Katakana	51 102
Kayah Li	138
Kharoshthi	1017
Khmer	88 108
Lao	26
Latin	1 2 3 4 30 130 131
Lepcha	134
Letter	36 55 1039 1040
Limbu	106
Linear B ideograms	1008
Linear B syllabary	1007
Lisu	150
Lycian	1024
Lydian	1026
Malayalam	24
Mandaic	157
Mathematical alphanumeric symbols	1006
Mathematical operators	39 101
Mathematical symbols	97 100
Meetei Mayek	156
MES	281 282
Mongolian	89
Months	55
Musical notation	1018
Musical symbols	1004 1005
Myanmar	87 90 155
New Tai Lue	114
Nko	128
Number	37 152 1009 1015
Ogham	82
Ol Chiki	135
Old Italic	1001
Old Persian	1016
Old South Arabian	1033
Old Turkic	1036
Optical character recognition	42
Oriya	20
Osmanya	1012
Phags-pa	132
Phaistos Disc	1023
Phoenician	1019
Phonetic extensions	109 116
Presentation forms	63 64 68 104 105
Private use	61 401
Punctuation	32 49 123
Radicals	77 78 79
Rejang	139
Rumi numeral symbols	1037
Runic	83
Samaritan	147
Saurashtra	137
Shape, shaping	205 206
Shavian	1011
Sinhala	84
Small form	67
Spacing modifier	6 125
Specials	70

Strokes	124	Technical	40
Subscripts, superscripts	33	Telugu	22
Sundanese	133	Thaana	86
Syllables, syllabics	71 74 76	Thai	25
Syloti Nagri	126	Tibetan	72 91
Symbols	9 34 35 36 47 49 97 100 1027	Tifinagh	121
		Ugaritic	1010
Syriac	85	Unicode	303 304 305 306 307 308 309 310 10646
Tagalog	93		
Tagbanwa	96	Vai	136
Tags	3001	Variation selectors	103 3003
Tai Tham	143	Vedic	149
Tai Viet	145	Vertical form	127
Tai Xuan Jing symbols	1014	Yi	76 77
Tail Le	107	Yijing hexagram symbols	111
Tamil	21	Zero-width	200
Tangut	1042		

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	TAGALOG	1700-171F
LATIN-1 SUPPLEMENT	00A0-00FF	HANUNOO	1720-173F
LATIN EXTENDED-A	0100-017F	BUHID	1740-175F
LATIN EXTENDED-B	0180-024F	TAGBANWA	1760-177F
IPA EXTENSIONS	0250-02AF	KHMER	1780-17FF
SPACING MODIFIER LETTERS	02B0-02FF	MONGOLIAN	1800-18AF
COMBINING DIACRITICAL MARKS	0300-036F	UNIFIED CANADIAN ABORIGINAL SYLLABICS EXTENDED	18B0-18FF
GREEK AND COPTIC	0370-03FF	LIMBU	1900-194F
CYRILLIC	0400-04FF	TAI LE	1950-197F
CYRILLIC SUPPLEMENT	0500-052F	NEW TAI LUE (Xishuang Banna Dai)	1980-19DF
ARMENIAN	0530-058F	KHMER SYMBOLS	19E0-19FF
HEBREW	0590-05FF	BUGINESE	1A00-1A1F
ARABIC	0600-06FF	TAI THAM	1A20-1AAF
SYRIAC	0700-074F	BALINESE	1B00-1B7F
ARABIC SUPPLEMENT	0750-077F	SUNDANESE	1B80-1BBF
THAANA	0780-07BF	BATAK	1BC0-1BFF
NKO	07C0-07FF	LEPCHA	1C00-1C4F
SAMARITAN	0800-083F	OL CHIKI	1C50-1C7F
MANDAIC	0840-085F	VEDIC EXTENSIONS	1CD0-1CFF
DEVANAGARI	0900-097F	PHONETIC EXTENSIONS	1D00-1D7F
BENGALI	0980-09FF	PHONETIC EXTENSIONS SUPPLEMENT	1D80-1DBF
GURMUKHI	0A00-0A7F	COMBINING DIACRITICAL MARKS SUPPLEMENT	1DC0-1DFF
GUJARATI	0A80-0AFF	LATIN EXTENDED ADDITIONAL	1E00-1EFF
ORIYA	0B00-0B7F	GREEK EXTENDED	1F00-1FFF
TAMIL	0B80-0BFF	GENERAL PUNCTUATION	2000-206F
TELUGU	0C00-0C7F	SUPERSCRIPTS AND SUBSCRIPTS	2070-209F
KANNADA	0C80-0CFF	CURRENCY SYMBOLS	20A0-20CF
MALAYALAM	0D00-0D7F	COMBINING DIACRITICAL MARKS FOR SYMBOLS	20D0-20FF
SINHALA	0D80-0DFF	LETTERLIKE SYMBOLS	2100-214F
THAI	0E00-0E7F	NUMBER FORMS	2150-218F
LAO	0E80-0EFF	ARROWS	2190-21FF
TIBETAN	0F00-0FFF	MATHEMATICAL OPERATORS	2200-22FF
MYANMAR	1000-109F	MISCELLANEOUS TECHNICAL	2300-23FF
GEORGIAN	10A0-10FF	CONTROL PICTURES	2400-243F
HANGUL JAMO	1100-11FF	OPTICAL CHARACTER RECOGNITION	2440-245F
ETHIOPIC	1200-137F	ENCLOSED ALPHANUMERICS	2460-24FF
ETHIOPIC SUPPLEMENT	1380-139F	BOX DRAWING	2500-257F
CHEROKEE	13A0-13FF	BLOCK ELEMENTS	2580-259F
UNIFIED CANADIAN ABORIGINAL SYLLABICS	1400-167F	GEOMETRIC SHAPES	25A0-25FF
OGHAM	1680-169F	MISCELLANEOUS SYMBOLS	2600-26FF
RUNIC	16A0-16FF		

DINGBATS	2700-27BF	CJK UNIFIED IDEOGRAPHS	4E00-9FFF
MISCELLANEOUS MATHEMATICAL SYMBOLS-A	27C0-27EF	YI SYLLABLES	A000-A48F
SUPPLEMENTAL ARROWS-A	27F0-27FF	YI RADICALS	A490-A4CF
BRAILLE PATTERNS	2800-28FF	LISU	A4D0-A4FF
SUPPLEMENTAL ARROWS-B	2900-297F	VAI	A500-A63F
MISCELLANEOUS MATHEMATICAL SYMBOLS-B	2980-29FF	CYRILLIC EXTENDED-B	A640-A69F
SUPPLEMENTAL MATHEMATICAL OPERATORS	2A00-2AFF	BAMUM	A6A0-A6FF
MISCELLANEOUS SYMBOLS AND ARROWS	2B00-2BFF	MODIFIER TONE LETTERS	A700-A71F
GLAGOLITIC	2C00-2C5F	LATIN EXTENDED-D	A720-A7FF
LATIN EXTENDED-C	2C60-2C7F	SYLOTI NAGRI	A800-A82F
COPTIC	2C80-2CFF	COMMON INDIC NUMBER FORMS	A830-A83F
GEORGIAN SUPPLEMENT	2D00-2D2F	PHAGS-PA	A840-A87F
TIFINAGH	2D30-2D7F	SAURASHTRA	A880-A8DF
ETHIOPIC EXTENDED	2D80-2DDF	DEVANAGARI EXTENDED	A8E0-A8FF
CYRILLIC EXTENDED-A	2DE0-2DFF	KAYAH LI	A900-A92F
SUPPLEMENTAL PUNCTUATION	2E00-2E7F	REJANG	A930-A95F
CJK RADICALS SUPPLEMENT	2E80-2EFF	HANGUL JAMO EXTENDED-A	A960-A97F
KANGXI RADICALS	2F00-2FDF	JAVANESE	A980-A9DF
IDEOGRAPHIC DESCRIPTION CHARACTERS	2FF0-2FFF	CHAM	AA00-AA5F
CJK SYMBOLS AND PUNCTUATION	3000-303F	MYANMAR EXTENDED-A	AA60-AA7F
HIRAGANA	3040-309F	TAI VIET	AA80-AADF
KATAKANA	30A0-30FF	MEETEI MAYEK	ABC0-ABFF
BOPOMOFO	3100-312F	HANGUL SYLLABLES	AC00-D7A3
HANGUL COMPATIBILITY JAMO	3130-318F	HANGUL JAMO EXTENDED-B	D7B0-D7FF
KANBUN (CJK miscellaneous)	3190-319F	PRIVATE USE AREA	E000-F8FF
BOPOMOFO EXTENDED	31A0-31BF	CJK COMPATIBILITY IDEOGRAPHS	F900-FAFF
CJK STROKES	31C0-31EF	ALPHABETIC PRESENTATION FORMS	FB00-FB4F
KATAKANA PHONETIC EXTENSIONS	31F0-31FF	ARABIC PRESENTATION FORMS-A	FB50-FDFF
ENCLOSED CJK LETTERS AND MONTHS	3200-32FF	VARIATION SELECTORS	FE00-FE0F
CJK COMPATIBILITY	3300-33FF	VERTICAL FORMS	FE10-FE1F
CJK UNIFIED IDEOGRAPHS EXTENSION A	3400-4DBF	COMBINING HALF MARKS	FE20-FE2F
YIJING HEXAGRAM SYMBOLS	4DC0-4DFF	CJK COMPATIBILITY FORMS	FE30-FE4F
		SMALL FORM VARIANTS	FE50-FE6F
		ARABIC PRESENTATION FORMS-B	FE70-FEFE
		HALFWIDTH AND FULLWIDTH FORMS	FF00-FFEF
		SPECIALS	FFF0-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 SOUTH ARABIAN	10A60-10A7F
LINEAR B IDEOGRAMS	10080-100FF	AVESTAN	10B00-10B3F
AEGEAN NUMBERS	10100-1013F	INSCRIPTIONAL PARTHIAN	10B40-10B5F
ANCIENT GREEK NUMBERS	10140-1018F	INSCRIPTIONAL PAHLAVI	10B60-10B7F
ANCIENT SYMBOLS	10190-101CF	OLD TURKIC	10C00-10C4F
PHAISTOS DISC	101D0-101FF	RUMI NUMERAL SYMBOLS	10E60-10E7F
LYCIAN	10280-1029F	BRAHMI	11000-1107F
CARIAN	102A0-102DF	KAITHI	11080-110CF
OLD ITALIC	10300-1032F	CUNEIFORM	12000-123FF
GOTHIC	10330-1034F	CUNEIFORM NUMBERS AND PUNCTUATION	12400-1247F
UGARITIC	10380-1039F	EGYPTIAN HIEROGLYPHS	13000-1342F
OLD PERSIAN	103A0-103DF	TANGUT	17000-1871F
DESERET	10400-1044F	BYZANTINE MUSICAL SYMBOLS	1D000-1D0FF
SHAVIAN	10450-1047F	MUSICAL SYMBOLS	1D100-1D1FF
OSMANYA	10480-104AF	ANCIENT GREEK MUSICAL NOTATION	1D200-1D24F
CYPRIOT SYLLABARY	10800-1083F	TAI XUAN JING SYMBOLS	1D300-1D35F
IMPERIAL ARAMAIC	10840-1085F	COUNTING ROD NUMERALS	1D360-1D37F
PHOENICIAN	10900-1091F	MATHEMATICAL ALPHANUMERIC SYMBOLS	1D400-1D7FF
LYDIAN	10920-1093F	MAHJONG TILES	1F000-1F02F
KHAROSHTHI	10A00-10A5F		

DOMINO TILES	1F030-1F09F	ENCLOSED IDEOGRAPHIC SUPPLEMENT	1F200-1F2FF
ENCLOSED ALPHANUMERIC SUPPLEMENT	1F100-1F1FF		

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 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)

The following fixed collections (see 4.25) contain the whole UCS assigned character content as it was when they were created. The Unicode collections are described in A.1.

A.3.1 301 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
01	00-F5 FA-FF
02	00-17 50-A8 B0-DE E0-E9
03	00-45 60-61 74-75 7A 7E 84-8A 8C 8E-A1 A3-CE D0-D6 DA DC DE E0 E2-F3
04	01-0C 0E-4F 51-5C 5E-86 90-C4 C7-C8 CB-CC D0-EB EE-F5 F8-F9
05	31-56 59-5F 61-87 89 91-A1 A3-B9 BB-C4 D0-EA F0-F4
06	0C 1B 1F 21-3A 40-52 60-6D 70-B7 BA-BE C0-CE D0-ED F0-F9
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
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
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
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
0D	02-03 05-0C 0E-10 12-28 2A-39 3E-43 46-48 4A-4D 57 60-61 66-6F
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
0F	00-47 49-69 71-8B 90-95 97 99-AD B1-B7 B9
10	A0-C5 D0-F6 FB
11	00-59 5F-A2 A8-F9
1E	00-9B A0-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
20	00-2E 30-46 6A-70 74-8E A0-AB D0-E1
21	00-38 53-82 90-EA
22	00-F1
23	00 02-7A
24	00-24 40-4A 60-EA
25	00-95 A0-EF
26	00-13 1A-6F
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.3.2 299 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 IO/IEC 10646-1.

A.3.3 302 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
01	00-FF
02	00-1F 22-33 50-AD B0-EE
03	00-4E 60-62 74-75 7A 7E 84-8A 8C 8E-A1 A3-CE D0-D7 DA-F3
04	00-86 88-89 8C-C4 C7-C8 CB-CC D0-F5 F8-F9
05	31-56 59-5F 61-87 89-8A 91-A1 A3-B9 BB-C4 D0-EA F0-F4
06	0C 1B 1F 21-3A 40-55 60-6D 70-ED F0-FE
07	00-0D 0F-2C 30-4A 80-B0
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
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
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
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
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
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
0F	00-47 49-6A 71-8B 90-97 99-BC BE-CC CF
10	00-21 23-27 29-2A 2C-32 36-39 40-59 A0-C5 D0-F6 FB
11	00-59 5F-A2 A8-F9
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
13	00-0E 10 12-15 18-1E 20-46 48-5A 61-7C A0-F4
14-15	1401-15FF
16	00-76 80-9C A0-F0
17	80-DC E0-E9
18	00-0E 10-19 20-77 80-A9
1E	00-9B A0-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
20	00-46 48-4D 6A-70 74-8E A0-AF D0-E3
21	00-3A 53-83 90-F3
22	00-F1
23	00-7B 7D-9A
24	00-26 40-4A 60-EA

25	00-95 A0-F7
26	00-13 19-71
27	01-04 06-09 0C-27 29-4B 4D 4F-52 56 58-5E 61-67 76-94 98-AF B1-BE
28	00-FF
2E	80-99 9B-F3
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.3.4 340 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 10646:2003 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
103	VARIATION SELECTORS
108	KHMER SYMBOLS
111	YIJING HEXAGRAM SYMBOLS

Row	Values within row
02	20-21 34-36 AE-AF EF-FF
03	4F-57 5D-5F 63-6F D8-D9 F4-FB
04	8A-8B C5-C6 C9-CA CD-CE
05	00-0F
06	00-03 0D-15 56-58 6E-6F EE-EF FF
07	2D-2F 4D-4F B1
09	04 BD
0A	01 03 8C E1-E3 F1
0B	35 71 F3-FA
0C	BC-BD
10	F7-F8
17	00-0C 0E-14 20-36 40-53 60-6C 6E-70 72-73 DD F0-F9
19	00-1C 20-2B 30-3B 40 44-4F 50-6D 70-74
1D	00-6B
20	47 4E-54 57 5F-63 71 B0-B1 E4-EA
21	3B 3D-4B F4-FF
22	F2-FF
23	7C 9B-D0
24	EB-FF
25	96-9F F8-FF
26	14-17 72-7D 80-91 A0-A1
27	68-75 D0-EB
2B	00-0D
30	3B-3D 95-96 9F-A0 FF
32	1D-1E 50-5F 7C-7D B1-BF CC-CF
33	77-7A DE-DF FF
A4	A2-A3 B4 C1 C5
FA	30-6A
FD	FC-FD
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 (9810 characters) 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 LINE FEED as end of line mark, that specifies, after a 11-lines header, as many lines as IICORE characters; each containing the following information in fixed length field.

- 1st field: BMP or SIP code point (`0hhhh`), (`2hhhh`), normative.
- 2nd field: Hanzi G usage identifier (`G0a`), (`G1a`), (`G3a`), (`G5a`), (`G7a`), (`G8a`), (`G9a`), or (`GEa`), informative.
- 3rd field: Hanzi T usage identifier (`T1a`), (`T2a`), (`T3a`), (`T4a`), (`T5a`) or (`TFa`), informative.
- 4th field: Kanji J usage identifier (`J1A`), in-formative.
- 5th field: Hanzi H usage identifier (`H1a`), in-formative.
- 6th field: Hanja K usage identifier (`K0a`), (`K1a`), (`K2a`) or (`K3a`), informative.
- 7th field: Hanzi M (for Macao SAR) usage identifier (`M1a`), informative.
- 8th field: Hanja KP usage identifier (`P0a`), informative.
- 9th field: General category, informative (`A`, `B` or `C` in decreasing order of priority).

The format definition uses ‘h’ as a hexadecimal unit and ‘a’ as an enumerated unit for letters from ‘A’ to ‘G’. Uppercase characters and digits between parentheses appear as shown.

NOTE 2 – The usage information provided in this subclause describes the usage and priority level of individual IICORE characters in the context of each source (G, T, J, H, K, M, and KP). This should not be confused with the source references for CJK Ideographs in 23 which establish the identity of all CJK Ideographs.

Click on this highlighted text to access the reference file.

NOTE 3 – The content is also available as a separate viewable file in the same file directory as this document. The file is named: “IICORE.txt”.

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 (3695 characters) 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 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: "JIEx.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 5801 characters.

NOTE – 2742 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

The collections specified within this clause 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.5.1 281 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
26	6A

A.5.2 282 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.5.3 283 MODERN EUROPEAN SCRIPTS

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

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

A.5.4 284 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 Indentifiers

<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.5.5 285 BASIC JAPANESE

The fixed collection 285 BASIC JAPANESE is a core Japanese subset. Its 6884 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
03	91-A1 A3-A9 B1-C1 C3-C9
04	01 10-4F 51
20	10 14 16 18-19 1C-1D 20-21 25-26 30 32-33 3B 3E
21	03 2B 90-93 D2 D4
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
23	12

25	00-03 0C 0F-10 13-14 17-18 1B-1D 20 23-25 28 2B-2C 2F-30 33-34 37-38 3B-3C 3F 42 4B A0-A1 B2-B3 BC-BD C6-C7 CB CE-CF EF
26	05-06 40 42 6A 6D 6F
30	00-03 05-15 1C 41-93 9B-9E A1-F6 FB-FE

A.5.6 286 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
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
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
03	00-04 06 08 0B-0C 0F 18-1A 1C-20 24-25 29-2A 2C 2F-30 34 39-3D 61 C2
1E	3E-3F
1F	70-73
20	13 22 3C 3F 42 47-49 51 AC
21	0F 13 16 21 27 35 53-55 60-6B 70-7B 94 96-99 C4 E6-E9
22	05 09 13 1F 25-26 2E 43 45 48 62 76-77 84-85 8A-8B 95-97 BF DA-DB
23	05-06 18 BE-CC CE
24	23 60-73 D0-E9 EB-FE
25	B1 B6-B7 C0-C1 C9 D0-D3 E6
26	00-03 0E 16-17 1E 60-69 6B-6C 6E
27	13 56 76-7F
29	34-35 BF FA-FB
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 51 57 7B-7E 8E-8F 9C-9E A1 C4 CB CD
FE	45-46
FF	5F-60

A.5.7 287 COMMON JAPANESE

The fixed collection 287 COMMON JAPANESE is a core Japanese subset containing 7493 characters. It includes a fixed collection from A.5 and several ranges of code points.

Planes 00-10

Collection number and name	
285	BASIC JAPANESE

Plane 00

Row	Values within row
20	15
21	16 21 60-69 70-79
22	11 1F 25 2E BF
24	60-73
30	1D 1F
32	31-32 39 A4-A8
33	03 0D 14 18 22-23 26-27 2B 36 3B 49-4A 4D 51 57 7B-7E 8E-8F 9C-9E A1 C4 CD
4E	28 E1 FC
4F	00 03 39 56 8A 92 94 9A C9 CD FF
50	1E 22 40 42 46 70 94 D8 F4
51	4A 64 9D BE EC
52	15 9C A6 AF C0 DB
53	00 07 24 72 93 B2 DD
54	8A 9C A9 FF
55	86
57	59 65 AC C7-C8
58	9E B2
59	0B 53 5B 5D 63 A4 BA
5B	56 C0 D8 EC
5C	1E A6 BA F5
5D	27 42 53 6D B8-B9 D0
5F	21 34 45 67 B7 DE
60	5D 85 8A D5 DE F2
61	11 20 30 37 98
62	13 A6
63	F5
64	60 9D CE
65	4E
66	00 09 15 1E 24 2E 31 3B 57 59 65 73 99 A0 B2 BF FA-FB
67	0E 66 BB C0
68	01 44 52 C8 CF
69	68 98 E2
6A	30 46 6B 73 7E E2 E4
6B	D6
6C	3F 5C 6F 86 DA
6D	04 6F 87 96 AC CF F2 F8 FC
6E	27 39 3C 5C BF
6F	88 B5 F5
70	05 07 28 85 AB BB
71	04 0F 46-47 5C C1 FE
72	B1 BE
73	24 77 BD C9 D2 D6 E3 F5
74	07 26 29-2A 2E 62 89 9F
75	01 2F 6F
76	82 9B-9C 9E A6
77	46
78	21 4E 64 7A
79	30 94 9B

7A	D1 E7 EB
7B	9E
7D	48 5C A0 B7 D6
7E	52 8A
7F	47 A1
83	01 62 7F C7 F6
84	48 B4 DC
85	53 59 6B B0
88	07 F5
89	1C
8A	12 37 79 A7 BE DF F6
8B	53 7F
8C	F0 F4
8D	12 76
8E	CF
90	67 DE
91	15 27 D7 DA DE E4-E5 ED-EE
92	06 0A 10 39-3A 3C 40 4E 51 59 67 77-78 88 A7 D0 D3 D5 D7 D9 E0 E7 F9 FB FF
93	02 1D-1E 21 25 48 57 70 A4 C6 DE F8
94	31 45 48
95	92
96	9D AF
97	33 3B 43 4D 4F 51 55
98	57 65
99	27 9E
9A	4E D9 DC
9B	72 75 8F B1 BB
9C	00
9D	6B 70
9E	19 D1
F9	29 DC
FA	0E-2D
FF	01-5E 61-9F E0-E5

A.6 Unicode collections

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.6.1 303 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.6.2 304 UNICODE 3.2

The fixed collection 304 UNICODE 3.2 consists of fixed collections from A.1 and 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

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

A.6.3 305 UNICODE 4.0

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

A.6.4 306 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
03	58-5C FC-FF
04	F6-F7
05	A2 C5-C7
06	0B 1E 59-5E
07	50-6D
09	7D CE
0B	B6 E6
0F	D0-D1
10	F9-FA FC
12	07 47 87 AF CF EF
13	0F 1F 47 5F-60 80-99
19	80-A9 B0-C9 D0-D9 DE-DF
1A	00-1B 1E-1F
1D	6C-C3
20	55-56 58-5E 90-94 B2-B5 EB
21	3C 4C
23	D1-DB
26	18 7E-7F 92-9C A2-B1
27	C0-C6
2B	0E-13
2C	00-2E 30-5E 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
2E	00-17 1C-1D
31	C0-CF
32	7E
9F	A6-BB
A7	00-16
A8	00-2B
FA	70-D9
FE	10-19

Plane 01

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

A.6.5 307 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
03	7B-7D
04	CF FA-FF
05	10-13 BA
07	C0-FA
09	7B-7C 7E-7F
0C	E2-E3 F1-F2
1B	00-4B 50-7C
1D	C4-CA FE-FF
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.6.6 308 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-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

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

Plane 10

Row	Values within row
00-FF	0000-FFFD

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
04	87
05	14-23
06	06-0A 16-1A 3B-3F
07	6E-7F
09	71-72
0A	51 75
0B	44 62-63 D0
0C	3D 58-59 62-63 78-7F
0D	3D 44 62-63 70-75 79-7F
0F	6B-6C CE D2-D4
10	22 28 2B 33-35 3A-3F 5A-8A
18	AA
1B	80-AA AE-B9
1C	00-37 3B-49 4D-7F
1D	CB-E6
1E	9C-9F FA-FF

20	64 F0
21	4F 85-88
26	9D B3-BC C0-C3
27	CC EC-EF
2B	1B-1F 24-4C 50-54
2C	6D-6F 71-73 78-7D
2D	E0-FF
2E	18-1B 1E-30
31	2D D0-E3
9F	BC-C3
A5	00-FF
A6	00-2B 40-5F 62-73 7C-97
A7	1B-1F 22-8C FB-FF
A8	80-C4 CE-D9
A9	00-53 5F
AA	00-36 40-4D 50-59 5C-5F
FE	24-26

Plane 01

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

A.6.7 309 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
01-02	0100-02FF
03	00-77 7A-7E 84-8A 8C 8E-A1 A3-FF
04	00-FF
05	00-25 31-56 59-5F 61-87 89-8A 91-C7 D0-EA F0-F4
06	00-03 06-1B 1E-5E 60-FF
07	00-0D 0F-4A 4D-B1 C0-FA
08	00-2D 30-3E
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
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-D8
10	00-C5 D0-FC
11	00-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
13	00-10 12-15 18-5A 5F-7C 80-99 A0-F4
14-15	1400-15FF
16	00-9C 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 E0-FC
19	00-1C 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
1B	00-4B 50-7C 80-AA AE-B9
1C	00-37 3B-49 4D-7F D0-F2
1D	00-E6 FD-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-B8 D0-F0
21	00-89 90-FF
22	00-FF
23	00-E8
24	00-26 40-4A 60-FF
25	00-FF
26	00-FF
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 60-7F 80-F1 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	00-31 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-FF
33	00-FF
34-4C	3400-4CFF
4D	00-B5 C0-FF
4E-9F	4E00-9FC6
A0-A3	A000-A3FF
A4	00-8C 90-C6 D0-FF
A5	00-FF
A6	00-2B 40-5F 62-73 7C-97 A0-F7
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-7A 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
01	00-02 07-33 37-8A 90-9B D0-FD
02	80-9C A0-D0
03	00-1E 20-23 30-4A 80-9D 9F-C3 C8-D5
04	00-9D A0-A9
08	00-05 08 0A-35 37-38 3C 3F-55 57-5F
09	00-1B 1F-39 3F
0A	00-03 05-06 0C-13 15-17 19-33 38-3A 3F-47 50-58 60-7F
0B	00-35 39-3F-55 58-72 78-7F
0C	00-48
0E	60-7E
10	80-BD C0-C1
20-22	2000-22FF
23	00-6E
24	00-62 70-73
30-34	3000-342E
D0	00-F5
D1	00-26 29-DD
D2	00-45
D3	00-56 60-71
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-A5 A8-FF
D7	00-CB CE-FF
F0	00-2B 30-93
F1	00-0A 10-2D 31 3D 3F 42 46 4A-4F 57 5F 79 7B-7C 7F 8A-8D
F2	00 10-30 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-FFFD

Plane 10

Row	Values within row
00-FF	0000-FFFD

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
08	00-2D 30-3E
09	00 4E 55 79-7A FB
0F	D5-D8
10	9A-9D
11	5A-5E A3-A7 FA-FF
14	00
16	77-7F
18	E0-FC
19	AA-AB DA
1A	20-5E 60-7C 7F-89 90-99 A0-AD
1C	D0-F2
1D	FD
20	B6-B8
21	50-52 89
23	E8
26	9E-9F BD-BF C4-FF
2C	70 7E-7F EB-F1
2D	E0-FF
2E	31
32	44-4F FF
9F	C4-C6
A4	D0-FF
A6	A0-F7
A8	30-39 E0-FB
A9	60-7C 80-CD CF-D9 DE-DF
AA	60-7A 80-C2 DB-DF

AB	C0-ED F0-F9
D7	B0-C6 CB-FB
FA	6B-6D

Plane 01

Row	Values within row
08	40-55 57-5F
09	1A-1B
0A	60-7F
0B	00-35 39-55 58-72 78-7F
0C	00-48
0E	60-7E
10	80-BD C0-C1
30-34	3000-342E
F1	00-0A 10-2D 31 3D 3F 42 46 4A-4F 57 5F 79 7B-7C 7F 8A-8D
F2	00 10-30 40-48

Plane 02

Row	Values within row
A7-B7	A700-B734

A.6.8 310 UNICODE 6.0

The fixed collection 310 UNICODE 6.0 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	00-27 31-56 59-5F 61-87 89-8A 91-C7 D0-EA F0-F4
06	00-03 06-1B 1E-5E 60-FF
07	00-0D 0F-4A 4D-B1 C0-FA
08	00-2D 30-3E 40-5B 5E
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
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 66-77 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
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 12-3A 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-D8
10	00-C5 D0-FC
11	00-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
13	00-10 12-15 18-5A 5F-7C 80-99 A0-F4
14-15	1400-15FF
16	00-9C 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 B0-D1 D3-D7 D9-DA E0-FC
19	00-1C 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
1B	00-4B 50-7C 80-AA AE-B9 C0-F3 FA-FF
1C	00-37 3B-49 4D-7F D0-F2
1D	00-E6 FD-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-B8 D0-F0
21	00-89 90-FF
22	00-FF
23	00-E8
24	00-26 40-4A 60-FF
25	00-FF
26	00-FF
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 60-7F 80-F1 F9-FF
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
2E	00-31 80-99 9B-F3
2F	00-D5 F0-FB
30	00-3F 41-97 99-FF
31	05-2D 31-8E 90-B7 C0-E3 F0-FF
32	00-1E 20-FF
33	00-FF
34-4C	3400-4CFF
4D	00-B5 C0-FF
4E-9F	4E00-9FC6
A0-A3	A000-A3FF
A4	00-8C 90-C6 D0-FF
A5	00-FF
A6	00-2B 40-5F 62-73 7C-97 A0-F7
A7	00-8E 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-7A 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-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
00	00-0B 0D-26 28-3A 3C-3D 3F-4D 50-5D 80-FA
01	00-02 07-33 37-8A 90-9B D0-FD
02	80-9C A0-D0
03	00-1E 20-23 30-4A 80-9D 9F-C3 C8-D5
04	00-9D A0-A9
08	00-05 08 0A-35 37-38 3C 3F-55 57-5F
09	00-1B 1F-39 3F
0A	00-03 05-06 0C-13 15-17 19-33 38-3A 3F-47 50-58 60-7F
0B	00-35 39-3F-55 58-72 78-7F
0C	00-48
0E	60-7E
10	00-4D 52-6F 80-BD C0-C1
20-22	2000-22FF
23	00-6E
24	00-62 70-73
30-34	3000-342E
70-87	7000-8715
D0	00-F5
D1	00-26 29-DD
D2	00-45
D3	00-56 60-71
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-A5 A8-FF
D7	00-CB CE-FF
F0	00-2B 30-93
F1	00-0A 10-2D 31 3D 3F 42 46 4A-4F 57 5F 79 7B-7C 7F 8A-8D
F2	00 10-30 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-FFFD

Plane 10

Row	Values within row
00-FF	0000-FFFD

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.

Plane 00-10

Collection number and name	
309	UNICODE 5.2

Plane 00

Row	Values within row
05	26-27
08	40-5B 5E
0B	72-77
0D	29 3A
18	BB-D1 D3-D7 D9-DA
1B	C0-F3 FA-FF
2D	70 7F
30	97
A7	8D-8E
FB	B2-C1

Plane 01

Row	Values within row
10	00-4D 52-6F
70-87	7000-8715

Annex B
(normative)
List of combining characters

NOTE – Replaced by formal character class definition, see 4.14

Annex C
(normative)
Transformation format for planes 1 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 CC-data element. 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 Zero-width boundary indicators

The following characters are used to indicate whether or not the adjacent characters are separated by a word boundary or hyphenation boundary. Each of these zero-width boundary indicators has no width in its usual own presentation.

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 CC-data-element to encode a possible hyphenation point (for example: "tug{00AD}gumi"), the character representation remains otherwise unchanged. When encoding a CC-data-element that includes characters encoding hard line breaks, including actual hyphenations, the character representation of the text sequence must reflect any changes due to hyphenation (for example: "tugg{2010}" / "gumi", where / represents the line break).

NOTE 2 – The notations {00AD} and {2010} indicate the inclusion of the corresponding code points: 00AD and 2010 into the CC-data-elements. The curly brackets "{}" are not part of the CC-data elements.

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 3 – For additional usages of the ZERO WIDTH NO-BREAK SPACE for "signature", see annex H.

The following characters 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.

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.1.2 Format separators

The following characters 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.1.3 Bidirectional text formatting

The following characters 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 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).

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

The following 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 the above characters.

F.2 Script-specific format characters

F.2.1 Symmetric swapping format characters

The following 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 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 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.2.4 Tifinagh consonant joiner

TIFINAGH CONSONANT JOINER (2D7E): This character suppresses inherent vowel, functions to indicate that the previous character and the following character are part of a bi-consonant cluster.

F.3 Interlinear annotation characters

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

The scope of these characters is the subsequent sequence of digits (plus certain other characters), with the exact specification as defined in the Unicode Standard, Version 5.0 (see Annex M for referencing information), for ARABIC END OF AYAH.

F.5 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.6 Language tagging using Tag characters

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.6.1 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 encoded using tag characters. No other characters are valid for expressing the tag arguments.

F.6.2 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 cc-data-element 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.6.3 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 must 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.6.4 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 LINE FEED as end of line mark, that specifies, after a 4-lines header, all the character names from ISO/IEC 10646 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 COMPATIBILITY IDEOGRAPHS, and

CJK COMPATIBILITY IDEOGRAPHS SUPPLEMENT).

The format of the file, after the header, is as follows:

01-05 octet: UCS-4 five-digit abbreviated form,

06 octet: TAB character,

07-end of line: 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 10.

Annex I
(informative)
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 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 ISO/IEC 10646.

I.1.1 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
- 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. (D_1 D_2) or (D_1 D_2 D_3).

NOTE 2 – An IDS is restricted to no more than 16 characters in length. Also no more than six ideographs and/or radicals may occur between any two instances of an IDC character within an IDS.

I.1.2 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 D_1 on the left and D_2 on the right.

IDEOGRAPHIC DESCRIPTION CHARACTER ABOVE TO BELOW (2FF1): The IDS introduced by this character describes the abstract form of the ideograph with D_1 above D_2.

IDEOGRAPHIC DESCRIPTION CHARACTER LEFT TO MIDDLE AND RIGHT (2FF2): The IDS introduced by this character describes the abstract form of the ideograph with D_1 on the left of D_2, and D_2 on the left of D_3.

IDEOGRAPHIC DESCRIPTION CHARACTER ABOVE TO MIDDLE AND BELOW (2FF3): The IDS introduced by this character describes the abstract form of the ideograph with D_1 above D_2, and D_2 above D_3.

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.

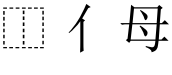
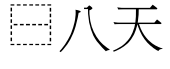
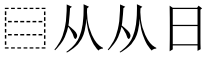
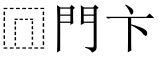
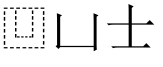
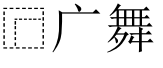
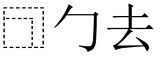
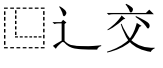
IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM UPPER LEFT (2FF8): The IDS introduced by this character describes the abstract form of the ideograph with D_1 at the top left corner of D_2, and partly surrounding D_2 above and to the left.

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 positions of DCs	Example of IDS	IDS example represents:
LEFT TO RIGHT	2	IDC-LTR D_1 D_2			毋
ABOVE TO BELOW	2	IDC-ATB D_1 D_2			矣
LEFT TO MIDDLE AND RIGHT	3	IDC-LMR D_1 D_2 D_3			衍
ABOVE TO MIDDLE AND BELOW	3	IDC-AMB D_1 D_2 D_3			咎
FULL SURROUND	2	IDC-FSD D_1 D_2			圓
SURROUND FROM ABOVE	2	IDC-SAV D_1 D_2			閑
SURROUND FROM BELOW	2	IDC-SBL D_1 D_2			凷
SURROUND FROM LEFT	2	IDC-SLT D_1 D_2			匿
SURROUND FROM UPPER LEFT	2	IDC-SUL D_1 D_2			廡
SURROUND FROM UPPER RIGHT	2	IDC-SUR D_1 D_2			匆
SURROUND FROM LOWER LEFT	2	IDC-SLL D_1 D_2			这
OVERLAID	2	IDC-OVL D_1 D_2	 *		巫

* NOTE – 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

This annex is applicable to a widely-used class of devices that can store received CC-data elements for subsequent retransmission.

This recommendation is intended to ensure that loss of information is minimized between the receipt of a CC-data-element 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 CC-data-elements 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 ISO/IEC 10646 except in clause 12 is different from other character coding standards such as ISO/IEC 2022, ISO/IEC 6429 and ISO 8859. This annex clarifies the relationship between the two notations.

In ISO/IEC 10646, 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 ISO/IEC 10646 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 weight 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 weight 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 ISO/IEC 10646 octet value notation 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. This annex 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 clause 24.5 and 24.6 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

Name:	Acronym
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 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 named UCS Sequence Identifiers 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 1 – 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 named UCS Sequence Identifiers, 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

Ộ	LATIN CAPITAL LETTER O WITH CIRCUMFLEX AND DOT BELOW
Ḉ	LATIN CAPITAL LETTER C WITH CEDILLA AND ACUTE
Ų́	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
A	GREEK CAPITAL LETTER ALPHA
A	CYRILLIC CAPITAL LETTER A

Guideline 8

Where possible, named UCS Sequence Identifiers are constructed by appending the names of the constituent elements together while eliding duplicate elements. Should this process result in a name that al-

ready exists, the name is modified suitably to guarantee uniqueness among character names and named UCS Sequence Identifiers. 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 followed in the list of characters of a particular standard by the name in parentheses which this character has in the script specified by this particular standard.

EXAMPLE

‿ UNDERTIE (Enotikon)

Annex M
(informative)
Sources of characters

Several sources and contributions were used for constructing this coded character set. In particular, characters of the following national and international standards are included in ISO/IEC 10646.

ISO 233:1984, *Documentation - Transliteration of Arabic characters into Latin characters*.

ISO/IEC 646:1991, *Information technology - ISO 7-bit coded character set for information interchange*.

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 5426:1983, *Extension of the Latin alphabet coded character set for bibliographic information interchange.*

ISO 5427:1984, *Extension of the Cyrillic alphabet coded character set for bibliographic information interchange.*

ISO 5428:1984, *Greek alphabet coded character set for bibliographic information interchange.*

ISO 6438:1983, *Documentation - African coded character set for bibliographic information interchange.*

ISO 6861, *Information and documentation - Glagolitic coded character set for bibliographic information interchange.*

ISO 6862, *Information and documentation - Mathematical coded character set for bibliographic information interchange.*

ISO 6937:1994, *Information technology - Coded graphic character sets 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 5: Latin/Cyrillic alphabet (1999)

Part 6: Latin/Arabic alphabet (1999)

Part 7: Latin/Greek alphabet

Part 8: Latin/Hebrew alphabet (1999)

Part 9: Latin alphabet No. 5 (1999)

Part 10: Latin alphabet No. 6 (1998).

ISO 8879:1986, *Information processing - Text and office systems - Standard Generalized Markup Language (SGML).*

ISO 8957:1996, *Information and documentation - Hebrew alphabet coded character sets for bibliographic information interchange.*

ISO 9036:1987, *Information processing - Arabic 7-bit coded character set for information interchange.*

ISO/IEC 9995-7:1994, *Information technology – Keyboard layouts for text and office systems – Part 7: Symbols used to represent functions.*

ISO/IEC 10367:1991, *Information technology - Standardized coded graphic character sets for use in 8-bit codes.*

ISO 10754:1984, *Information and documentation – Extension of the Cyrillic alphabet coded character set for non-Slavic languages for bibliographic information interchange.*

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.*

ISO/IEC TR 15285:1998, *Information technology - An operational model for characters and glyphs.*

ISO international register of character sets to be used with escape sequences. (registration procedure ISO 2375:1985) .

ANSI X3.4-1986 American National Standards Institute. *Coded character set - 7-bit American national standard code.*

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 Y10.20-1988 American National Standards Institute. *Mathematic signs and symbols for use in physical sciences and technology.*

ANSI Y14.5M-1982 American National Standard. *Engineering drawings and related document practices, dimensioning and tolerances.*

ANSI Z39.47-1985 American National Standards Institute. *Extended Latin alphabet coded character set for bibliographic use.*

ANSI Z39.64-1989 American National Standards Institute. *East Asian character code for bibliographic use.*

ASMO 449-1982 Arab Organization for Standardization and Metrology. *Data processing - 7-bit coded character set for information interchange.*

GB2312-80 *Code of Chinese Graphic Character Set for Information Interchange: Jishu Biaozhun Chubanshe* (Technical Standards Publishing).

NOTE – For additional sources of the CJK unified ideographs in ISO/IEC 10646 refer to clause 23.

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).

GBK (*Guo Biao Kuo*) *Han character internal code extension specification: Jishu Biaozhun Chubanshe* (Technical Standards Publishing, Beijing)

IS 13194:1991 Bureau of Indian Standards *Indian script code for information interchange - ISCII*

LTD 37(1610)-1988 *Indian standard code for information interchange.*

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.

JIS X 0201-1976 Japanese Standards Association. *Jouhou koukan you fugou (Code for Information Interchange).*

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 C 5601-1992 Korean Industrial Standards Association. *Jeongbo gyohwanyong buho (Code for Information Interchange).*

LVS 18-92 Latvian National Centre for Standardization and Metrology *Libiesu kodu tabula ar 191 simbolu.*

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.*

SLS 1134:1996 Sri Lanka Standards Institution *Sinhala character code for information interchange*.

TIS 620-2533 *Thai Industrial Standard for Thai Character Code for Computer.* (1990)

The following publications were also used as sources of characters for the Basic Multilingual Plane.

Allworth, Edward. *Nationalities of the Soviet East: Publications and Writing Systems*. New York, London, Columbia University Press, 1971. ISBN 0-231-03274-9.

Armbruster, Carl Hubert. *Initia Amharica: an Introduction to Spoken Amharic*. Cambridge, Cambridge University Press, 1908-20.

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

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

Beyer, Stephen V. *The classical Tibetan language*. State University of New York. ISBN 0-7914-1099-4

Bburx Ddie Su (= Bian Xiezhe). 1984. *Nuo-su bburma shep jie zzit: Syp-chuo se nuo bbur-ma syt mu*

curx su niep sha zho ddop ma bbur-ma syt mu wo yuop hop, Bburx Ddie da Su. [Chengdu]: Syp-chuo co cux tep yy ddurx dde. *Yi wen jian zi ben: Yi Han wen duizhao ban.* Chengdu: Sichuan minzu chubanshe. [*An examination of the fundamentals of the Yi script.* Chengdu: Sichuan National Press.]

Bburx Ddie Su. *Nip huo bbur-ma ssix jie: Nip huo bbur-ma ssi jie Bburx Ddie curx Su. = Yi Han zidian.* Chengdu: Sichuan minzu chubanshe, 1990. ISBN 7-5409-0128-4

Daniels, Peter T., and William Bright, eds. 1996. *The world's writing systems*. New York; Oxford: Oxford University Press. ISBN 0-19-507993-0

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.

Diringer, David. 1996. *The alphabet: a key to the history of mankind*. New Delhi: Munshiram Manoharlal. ISBN 81-215-0780-0

Esling, John. *Computer coding of the IPA: supplementary report*. Journal of the International Phonetic Association, 20:1 (1990), p. 22-26.

Faulmann, Carl. 1990 (1880). *Das Buch der Schrift*. Frankfurt am Main: Eichborn. ISBN 3-8218-1720-8

Friesen, Otto von. *Runorna*. Stockholm, A. Bonnier [1933]. (Nordisk kultur, 6).

Geiger, Wilhelm. *Maldivian Linguistic Studies*. New Delhi, Asian Educational Services, 1996. ISBN 81-206-1201-9.

Gunasekara, Abraham Mendis. 1986 (1891). *A comprehensive grammar of the Sinhalese language*. New Delhi: Asian Educational Services.

Haarmann, Harald. 1990. *Universalgeschichte der Schrift*. Frankfurt/Main; New York: Campus. ISBN 3-593-34346-0

Holmes, Ruth Bradley, and Betty Sharp Smith. 1976. *Beginning Cherokee: Talisgo galiquogi dideliquasdodi Tsalagi digoweli* . Norman: University of Oklahoma Press.

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.

Imprimerie Nationale. 1990. *Les caractères de l'Imprimerie Nationale*. Paris: Imprimerie Nationale Éditions. ISBN 2-11-081085-8

International Phonetic Association. *The International Phonetic Alphabet* (revised to 1989).

Jensen, Hans. 1969. *Die Schrift in Vergangenheit und Gegenwart*. 3., neubearbeitete und erweiterte Auflage. Berlin: VEB Deutscher Verlag der Wissenschaften.

Kefarnissy, Paul. *Grammaire de la langue araméenne syriaque*. Beyrouth, 1962.

Knuth, Donald E. *The TeXbook*. – 19th. printing, rev, – Reading, MA : Addison-Wesley, 1990.

Kuruch, Rimma Dmitrievna. *Saamsko-russkiy slovar'*. Moskva: Russkiy iazyk. 1985

Launhardt, Johannes. *Guide to Learning the Oromo (Galla) Language*. Addis Ababa, Launhardt [1973?]

Leslau, Wolf. *Amharic Textbook*. Weisbaden, Harrassowitz; Berkeley, University of California Press, 1968.

Mandarin Promotion Council, Ministry of Education, Taiwan. *Shiangtu yuyan biauyin fuhau shoutse* (*The Handbook of Taiwan Languages Phonetic Alphabet*). 1999.

Nakanishi, Akira. 1990. *Writing systems of the world: alphabets, syllabaries, pictograms*. Rutland, VT: Charles E. Tuttle. ISBN 0-8048-1654-9

Okell, John. 1971. *A guide to the romanization of Burmese*. (James G. Forlang Fund; 27) London: Royal Asiatic Society of Great Britain and Ireland.

Page, R. I. 1987. *Runes*. (Reading the Past; 4) Berkeley & Los Angeles: University of California Press. ISBN 0-520-06114-4

Pullum, Geoffrey K. *Phonetic symbol guide*. Geoffrey K. Pullum and William A. Ladusaw. – Chicago : University of Chicago Press, 1986.

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.

Roop, D. Haigh. 1972. *An introduction to the Burmese writing system*. New Haven and London: Yale University Press. ISBN 0-300-01528-3

Santos, Hector. 1994. *The Tagalog script.* (Ancient Philippine Scripts Series; 1). Los Angeles: Sushi Dog Graphics.

Santos, Hector. 1995. *The living scripts*. (Ancient Philippine Scripts Series; 2). Los Angeles: Sushi Dog Graphics.

Selby, Samuel M. *Standard mathematical tables.* – 16th ed. – Cleveland, OH : Chemical Rubber Co., 1968. Shepherd, Walter.

Shepherd, Walter. *Shepherd's glossary of graphic signs and symbols*. Compiled and classified for ready reference. – New York : Dover Publications, [1971].

Shinmura, Izuru. *Kojien – Dai 4-han*. – Tokyo : Iwanami Shoten, Heisei 3 [1991].

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

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 *Unicode Standard Annexes, UAX#9, The Unicode Bidirectional Algorithm, UAX#15 Unicode Normalization Forms, Version 4.0.0* 2003, and related Unicode Technical Reports, available at:
http://www.unicode.org/reports/

The following publications were also used as sources of characters for the Supplementary Multilingual Plane.

Deseret

Ivins, Stanley S. "The Deseret Alphabet" *Utah Humanities Review 1* (1947):223-39.

Old Italic

Bonfante, Larissa. 1996. "The scripts of Italy", in Peter T. Daniels and William Bright, eds. *The world's writing systems*. New York; Oxford: Oxford University Press. ISBN 0-19-507993-0

Gothic

Fairbanks, Sydney, and F. P. Magoun Jr. 1940. 'On writing and printing Gothic', in *Speculum* 15:313-16.

Byzantine Musical Symbols

ELOT 1373. *The Greek Byzantine Musical Notation System.* Athens, 1997 (ΣΕΠ ΕΛΟΤ 1373: 1997).

Musical Symbols

Heussenstamm, George. Norton Manual of Music Notation. New York: W. W. Norton, 1987

Rastall, Richard. Notation of Western Music: An Introduction. London: Dent, 1983

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 ISO/IEC 10646. 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 ISO/IEC 10646 a precise declaration of that repertoire should include the following parameters:

- identification of ISO/IEC 10646,
- the adopted subset of the repertoire, identified by one or more collection numbers,
- the CC-data-element 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 ISO/IEC 10646 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 ISO/IEC 10646 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 CC-data-element content definition, and is referred as ‘level-3 (3)’.

NOTE 2 – This version of the standard specifies a single definition for CC-data-element 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 ISO/IEC 10646. 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.

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 ISO/IEC 10646 is defined to be a "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 ISO/IEC 10646 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”.

Annex P
(informative)
Additional information on CJK Unified Ideographs

This annex contains additional information on CJK Unified Ideographs.

NOTE – The first edition of this 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 in this annex.

Each entry in this annex consists of the name of a character preceded by its code point, followed by the related additional information. Entries are arranged in ascending sequence of code point.

9FB9 CJK UNIFIED IDEOGRAPH-9FB9
9FBA CJK UNIFIED IDEOGRAPH-9FBA
9FBB CJK UNIFIED IDEOGRAPH-9FBB

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, 2099D, and 470C respectively.

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

This annex provides the full name and additional information of Hangul syllables through a linked file:

The content linked to is a plain text file, using ISO/IEC 646-IRV characters with 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 additional information 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 ISO/IEC 10646 are specified in 30. They are derived from many more ideographs which are found in various different national and regional standards for coded character sets (the "sources").

This annex 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 23.1.

Within the context of ISO/IEC 10646 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)

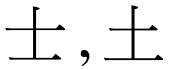
CJK COMPATIBILITY IDEOGRAPHS SUPPLEMENT (2F800-2FA1F).

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.

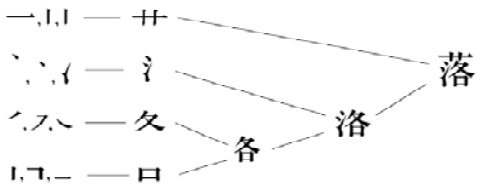


Figure S.1 - Component structure

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 most superior node, as shown in Figure S.2.

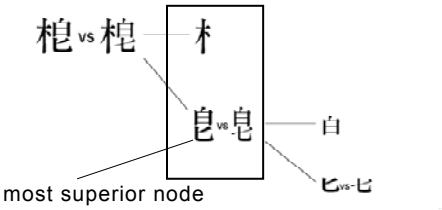


Figure S.2 - The most superior node of a component

The following features of each ideograph to be compared are examined:

a) the number of components,

b) the relative position of the components in each complete ideograph,

c) 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•厷, 降•夅

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

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.

拡•擴, 策•筞, ⺍•𤇾, 圣•巠, 㑒•僉, 区•區, 夹•夾,
単•單, 萑•雚, 戋•戔 賛•贊, 㐮•襄, 𢦏•韱, 間•閒,
朶•朵, 雋•隽, 恒•恆, 奐•奂, 从•从, 秦•秦, 夔•夔

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 three ideographs in each group below have different actual shapes, but they are considered to have the same abstract shape, and are therefore unified.

辶•辶•辶, 示•示•礻, 艮•艮•皀, 食•食•飠, 黄•黃, 𥁕•昷, 曷•曷,
包•包, 青•青, 毎•每, 冊•冊, 爭•争, 䍃•䍃•䍃, 彔•录,
步•歩, 者•者, 臭•臭, 幷•并, 骨•骨, 呂•吕, 直•直,
県•県, 吳•吴•吳, 眞•真•真, 爲•為, 単•单, 曾•曾•曾, 成•成,
專•専, 內•内, 晉•晋, 龜•龜, 艹•艹

The differences are further classified according to the following examples.

a) Differences in rotated strokes/dots

半•半, 勺•勺, 羽•羽, 酋•酋, 兼•兼, 益•益

b) Differences in overshoot at the stroke initiation and/or termination

身•身, 雪•雪, 拐•拐, 不•不, 非•非, 周•周

c) Differences in contact of strokes

奧•奥, 酉•酉, 查•査

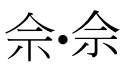
d) Differences in protrusion at the folded corner of strokes

巨•巨

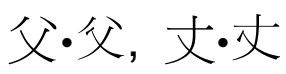
e) Differences in bent strokes

册•冊

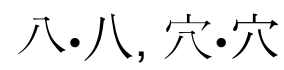
f) Differences in folding back at the stroke termination



g) Differences in accent at the stroke initiation



h) Differences in "rooftop" modification



i) Combinations of the above differences

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 30 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 C 5601-1989, KS C 5657-1991

NOTE – 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 this clause only applies to the CJK UNIFIED IDEOGRAPHS block specified in the Basic Multilingual Plane.

NOTE – 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 30 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	9th edition
2	Daikanwa Jiten 大漢和辞典	9th edition
3	Hanyu Dazidian 漢語大字典	1st edition
4	Daejaweon 大字源	1st 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

a) If an ideograph is found in the Kangxi Dictionary, it is positioned in the code chart in accordance with the Kangxi Dictionary order.

b) 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.

c) 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 code 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 at the beginning of this annex.

丟 丢	T	俱 俱	T	凈 淨	G	翍 翍	T
4E1F 4E22		4FF1 5036		51C0 51C8		524F 5259	
么 幺	GT	値 值	T	几 几	T	剝 剥	T
4E48 5E7A		5024 503C		51E2 51E3		525D 5265	
争 爭	GTJ	偷 偸	T	刃 刃	TJ	劒 劍	J
4E89 722D		5077 5078		5203 5204		5292 5294	
仞 仅	J	偽 僞	TJ	刊 刋	TJ	勻 匀	T
4EDE 4EED		507D 50DE		520A 520B		52FB 5300	
併 倂	T	兌 兑	T	刪 刪	T	単 単	T
4F75 5002		514C 5151		5220 522A		5355 5358	
侶 侶	T	兎 兔	TJ	別 別	T	即 卽	TK
4FA3 4FB6		514E 5154		5225 522B		5373 537D	
俣 俣	TJK	兗 兗	T	券 券	TJ	卷 卷	TJ
4FC1 4FE3		5156 5157		5238 52B5		5377 5DFB	
俞 俞	T	冊 冊	TJ	剎 剎	T	叁 参	GT
4FDE 516A		518A 518C		5239 524E		53C1 53C2	

參 叁 T 53C3 53C4	圖 圖 T 5716 5717	妍 妍 T 598D 59F8	寧 寧 T 5BDC 5BE7
吕 呂 T 5415 5442	巠 巠 T 5759 5DE0	姍 姍 T 59CD 59D7	寢 寢 GTJ 5BDD 5BE2
吞 吞 T 541E 5451	垮 垮 J 57D2 57D3	姬 姬 GT 59EB 59EC	專 專 J 5C02 5C08
吳 吴 吳 TJ 5433 5434 5449	墜 墜 T 5848 588D	娛 娛 娛 T 5A1B 5A2F 5A31	將 將 GTJ 5C06 5C07
吶 吶 T 5436 5450	塡 填 TJ 5861 586B	婕 婕 T 5A55 5AAB	尔 尔 T 5C13 5C14
告 告 T 543F 544A	增 增 T 5897 589E	媮 媮 T 5A7E 5AAE	尙 尚 T 5C19 5C1A
唧 唧 T 5527 559E	壯 壯 GTJ 58EE 58EF	媼 媼 TK 5AAA 5ABC	尷 尷 T 5C2A 5C2B
喻 喻 T 55A9 55BB	壽 壽 T 58FD 5900	嫣 嫣 T 5AAF 5B00	尷 尷 T 5C36 5C37
噓 噓 T 5618 5653	夐 夐 T 5910 657B	嫵 嫵 T 5B0E 5B14	屏 屏 T 5C4F 5C5B
嚏 嚏 GTJ 568F 5694	本 本 GTJ 5932 672C	嬤 嬤 GT 5B24 5B37	崢 崢 GT 5CE5 5D22
国 国 T 56EF 56FD	奧 奥 J 5965 5967	孳 孳 T 5B73 5B76	巔 巔 T 5DD3 5DD4
圈 圈 TJ 5708 570F	奬 獎 奬 TJ 5968 596C 734E	宮 宮 T 5BAB 5BAE	帡 帡 T 5E21 5E32
圓 圓 T 570E 5713	妝 妝 GT 5986 599D	寬 寬 T 5BDB 5BEC	帶 帶 TJ 5E2F 5E36

并 幷 T 5E76 5E77	恵 惠 TJ 6075 60E0	插 揷 揷 TJ 633F 63D2 63F7	曽 曾 J 66FD 66FE
廄 廏 T 5EC4 5ECF	悦 悦 T 6085 60A6	揑 揑 TJ 634F 63D1	栄 栄 T 67B4 67FA
弒 弑 T 5F11 5F12	悞 悞 T 609E 60AE	捜 搜 TJ 635C 641C	查 查 T 67E5 67FB
強 强 T 5F37 5F3A	悳 悳 T 60B3 60EA	掲 揭 T 63B2 63ED	柵 栅 T 67F5 6805
弾 弹 T 5F39 5F3E	愠 慍 T 6120 614D	揺 搖 揺 TJ 63FA 6416 6447	梲 棁 T 68B2 68C1
彐 彑 TJ 5F50 5F51	愼 慎 TJ 613C 614E	揾 搵 T 63FE 6435	榆 榆 T 6961 6986
彔 录 T 5F54 5F55	戯 戲 GT 6229 622C	撃 擊 TJ 6483 64CA	概 概 T 6982 69EA
彙 彙 T 5F59 5F5A	戲 戲 T 622F 6231	教 教 T 654E 6559	榲 榅 T 6985 69B2
彜 彝 J 5F5B 5F5C	戶 户 戶 T 6236 6237 6238	敓 敓 T 6553 655A	榖 榖 T 699D 6A27
彝 彝 T 5F5D 5F5E	戾 戾 T 623B 623E	既 既 T 65E2 65E3	槇 槙 J 69C7 69D9
彦 彦 T 5F65 5F66	抛 拋 T 629B 62CB	昂 昂 T 6602 663B	様 樣 TJ 69D8 6A23
徳 德 T 5FB3 5FB7	抜 拔 TJ 629C 62D4	晩 晚 T 665A 6669	横 横 T 6A2A 6A6B
徵 徵 T 5FB4 5FB5	挩 挩 T 6329 635D	暨 暨 T 66A8 66C1	歩 步 T 6B65 6B69

歲 歳	T	淸 清	T	瑤 瑶	TJ	簞 箪	T
6B72 6B73		6DF8 6E05		7464 7476		7BB3 7C08	
歿 殁	T	渴 渇	T	甁 瓶	T	簒 纂	T
6B7F 6B81		6E07 6E34		74F6 7501		7BE1 7C12	
殼 殻	GTJ	溫 温	T	產 産	T	粵 粤	T
6BBB 6BBC		6E29 6EAB		7522 7523		7CA4 7CB5	
毀 毁	T	溈 潙	T	瘦 痩	J	絕 絶	T
6BC0 6BC1		6E88 6F59		75E9 7626		7D55 7D76	
每 毎	T	溉 漑	T	皡 皞	T	綠 緑	T
6BCE 6BCF		6E89 6F11		76A1 76A5		7DA0 7DD1	
氳 氲	T	滾 滾	T	眞 真	TJ	緒 緒	T
6C32 6C33		6EDA 6EFE		771E 771F		7DD2 7DD6	
污 污	T	潛 潜	GTJK	眾 衆	TJK	緣 縁	T
6C5A 6C61		6F5B 6FF3		773E 8846		7DE3 7E01	
沒 没	TJ	瀨 瀬	T	研 研	T	緼 縕	T
6C92 6CA1		7028 702C		7814 784F		7DFC 7E15	
淨 浄	TJ	為 爲	GTJ	祿 禄	TJ	繈 繦	T
6D44 6DE8		70BA 7232		797F 7984		7E48 7E66	
涉 渉	T	煢 煢	GTJK	禿 禿	T	羹 羹	TJ
6D89 6E09		712D 7162		79BF 79C3		7FAE 7FB9	
淀 淀	T	熙 熙	J	稅 税	T	翺 翱	T
6D97 6D9A		7155 7199		7A05 7A0E		7FF6 7FFA	
淚 涙	T	煖 煖	T	穗 穂	TJ	胼 胼	T
6D99 6DDA		7174 7185		7A42 7A57		80FC 8141	
淶 渌	T	狀 状	GT	箏 箏	GJ	脫 脱	T
6DE5 6E0C		72B6 72C0		7B5D 7B8F		812B 8131	

腽 腽	T	蛻 蛻	T	达 达	T	閱 閱	T
817D 8183		86FB 8715		8FBE 8FD6		95B1 95B2	
舄 舄	GT	衛 衛	TJK	迸 迸	TJ	隉 隉	G
8203 8204		885B 885E		8FF8 902C		9667 9689	
舍 舍	TJ	衮 衮	TK	遙 遥	J	靑 青	T
820D 820E		886E 889E		9059 9065		9751 9752	
舖 舖	J	装 装	GJK	邢 邢	T	靜 靜	GTJ
8216 8217		88C5 88DD		90A2 90C9		9759 975C	
莊 莊	TJ	訐 訐	T	郎 郎	T	靭 靭	J
8358 838A		8A2E 8A7D		90CE 90DE		976D 9771	
菑 蓄	TJ	說 説	T	鄉 鄉 鄉	T	頹 頹	T
83D1 8458		8AAA 8AAC		90F7 9109 9115		9839 983D	
葢 葢	T	諫 諫	TJ	醞 醞	T	顏 顔	TJ
8480 8495		8ACC 8AEB		9196 919E		984F 9854	
蒋 蔣	GJ	謠 謡	J	醬 醤	J	顚 顛	J
848B 8523		8B20 8B21		91A4 91AC		985A 985B	
蔿 蔿	T	豺 豺	T	鈃 鈃	T	飲 飲	J
848D 853F		8C5C 8C63		9203 9292		98EE 98F2	
蕴 蕴	T	走 走	TJ	鋭 鋭	T	餅 餅	TJ
8570 8580		8D70 8D71		92B3 92ED		9905 9920	
薫 薫	T	軿 軿	T	錄 録	T	馱 駄	TJK
85AB 85B0		8EFF 8F27		9304 9332		99B1 99C4	
蘊 蘊	T	輜 輜	J	錬 錬	TK	駢 駢	TK
85F4 860A		8F1C 8F3A		932C 934A		99E2 9A08	
虛 虚	T	轀 轀	T	鎭 鎮	TJ	骯 骯	T
865A 865B		8F3C 8F40		93AD 93AE		9AA9 9AAB	

高 髙	T	鰛 鰮	TJ	鷆 鷆	J	黃 黃	T
9AD8 9AD9		9C1B 9C2E		9DC6 9DCF		9EC3 9EC4	
髮 髮	TJ	鳳 鳳	T	麵 麪	T	黑 黑	T
9AEA 9AEE		9CEF 9CF3		9EAA 9EAB		9ED1 9ED2	
鬪 鬪	T	鵝 鵝	J	麼 麼	T		
9B2C 9B2D		9D87 9DAB		9EBC 9EBD			

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 clause S.1.6.

冑 冑	non cognate	寶 寶	S.1.4.3	朐 朐	non cognate	稻 稻	S.1.4.3
5191 80C4		5BF3 5BF6		6710 80CA		7A32 7A3B	
冲 沖	S.1.4.3	廳 廳	S.1.4.1	朓 朓	non cognate	翱 翱	S.1.4.3
51B2 6C96		5EF0 5EF3		6713 8101		7FF1 7FF6	
决 決	S.1.4.3	懷 懷	S.1.4.1	朘 朘	non cognate	耆 耆 耆	S.1.4.3
51B3 6C7A		61D0 61F7		6718 8127		8007 8008 8009	
况 況	S.1.4.3	斀 斀	S.1.4.3	膧 膧	non cognate	聽 聽 聽	S.1.4.1
51B5 6CC1		6560 656A		6723 81A7		8074 807C 807D	
垛 垛	S.1.4.3	肦 肦	non cognate	朵 朵	S.1.4.3	荊 荊	S.1.4.2
579B 579C		670C 80A6		6735 6736		8346 834A	
擘 擘	S.1.4.2	朏 朏		灩 灩	S.1.4.3	躲 躲	S.1.4.3
5B7C 5B7D		non cognate 670F 80D0		7054 7067		8EB1 8EB2	

Annex T
(informative)
Language tagging using Tag Characters

NOTE – Moved to F.6.

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/.