INTERNATIONAL STANDARD

ISO/IEC 10646-1

Second edition
2000-09-15

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

Part 1: Architecture and Basic Multilingual Plane

iTeh STANDARD PREVIEW

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

Partie 1: Architecture et plan multilingue de base

(standards.iteh.ai)
ISO/IEC 10646-1:2000
https://standards.iteh.ai/catalog/standards/sist/23a1f63a-fe59-4266-aef0-8d25aead12e2/iso-iec-10646-1-2000

Reference number
ISO/IEC 10646-1:2000(E)

© ISO/IEC 2000

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.

iTeh STANDARD PREVIEW
(standards.iteh.ai)

ISO/IEC 10646-1:2000
https://standards.iteh.ai/catalog/standards/sist/23a1f63a-fe59-4266-aef0-8d25aead12e2/iso-iec-10646-1-2000

© ISO/IEC 2000

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

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

Printed in Switzerland

© ISO/IEC 2000 – All rights reserved

Contents

iTeh STANDARD PREVIEW
(standards.iteh.ai)
ISO/IEC 10646-1:2000
https://standards.iteh.ai/catalog/standards/sist/23a1ff3a-fe59-4266-aef0-8d25aead12e2/iso-iec-10646-1-2000

iTeh STANDARD PREVIEW
(standards.iteh.ai)
ISO/IEC 10646-1:2000
https://standards.iteh.ai/catalog/standards/sist/23a1f63a-fe59-4266-aef0-8d25aead12e2/iso-iec-10646-1-2000

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.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 3.

In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1. 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.

iTeh STANDARD PREVIEW

Attention is drawn to the possibility that some of the elements of this part 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.

(standards.iteh.ai)

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

ISO/IEC 10646-1:2000
https://standards.iteh.ai/catalog/standards/sist/23a1f63a-fe59-4266-aef0-8d25aead12e2/iso-iec-10646-1-2000

This second edition cancels and replaces the first edition (ISO/IEC 10646-1:1993), which has been technically revised. It also incorporates Amendments 1 to 13, 16 to 21 and 23, and Technical Corrigenda 1 and 2 to the first edition.

ISO/IEC 10646 consists of the following parts, under the general title *Information technology — Universal Multiple-Octet Coded Character Set (UCS)*:

— *Part 1: Architecture and Basic Multilingual Plane*

— *Part 2: Secondary Multilingual Plane for scripts and symbols, Supplementary Plane for CJK Ideographs, Special Purpose Plane*

Additional parts will specify other planes.

Annexes A to D form a normative part of this part of ISO/IEC 10646. Annexes E to S are for information only.

Introduction

ISO/IEC 10646 specifies the Universal Multiple-Octet Coded Character Set (UCS). It is applicable to the representation, transmission, interchange, processing, storage, input and presentation of the written form of the languages (scripts) of the world as well as additional symbols.

This part of ISO/IEC 10646 specifies the overall architecture and the Basic Multilingual Plane (BMP) of the UCS.

iTeh STANDARD PREVIEW
(standards.iteh.ai)
ISO/IEC 10646-1:2000
https://standards.iteh.ai/catalog/standards/sist/23a1f63a-fe59-4266-aef0-8d25aead12e2/iso-iec-10646-1-2000

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

Part 1: Architecture and Basic Multilingual Plane

1 Scope

ISO/IEC 10646 specifies the Universal Multiple-Octet Coded Character Set (UCS). It is applicable to the representation, transmission, interchange, processing, storage, input, and presentation of the written form of the languages of the world as well as of additional symbols.

This part of ISO/IEC 10646 specifies the overall architecture, and

- defines terms used in ISO/IEC 10646;

- describes the general structure of the coded character set;

- specifies the Basic Multilingual Plane (BMP) of the UCS, and defines a set of graphic characters used in scripts and the written form of languages on a worldwide scale;

- specifies the names for the graphic characters of the BMP, and their coded representations;

- specifies the four-octet (32-bit) canonical form of the UCS: UCS-4;

- specifies a two-octet (16-bit) BMP form of the UCS: UCS-2;

- specifies the coded representations for control functions;

- specifies the management of future additions to this coded character set.

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

NOTE 1 - The Unicode Standard, Version 3.0, provides a set of characters, names, and coded representations that are identical with those in this Part 1 of this International Standard. It additionallly provides details of character properties, processing algorithms, and definitions that are useful to implementors.

NOTE 2 - It is intended that character code positions for additional scripts and symbols will be allocated in this Part 1 of this International Standard when sufficient input and review is provided by national standards organizations or other qualified experts.

2 Conformance

2.1 General

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

2.2 Conformance of information interchange

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

a) all the coded representations of graphic characters within that CC-data-element conform to clauses 6 and 7, to an identified form chosen from clause 13 or annex C or annex D, and to an identified implementation level chosen from clause 14;

b) all the graphic characters represented within that CC-data-element are taken from those within an identified subset (clause 12);

c) all the coded representations of control functions within that CC-data-element conform to clause 15.

A claim of conformance shall identify the adopted form, the adopted implementation level and the adopted subset by means of a list of collections and/or characters.

2.3 Conformance of devices

A device is in conformance with ISO/IEC 10646 if it conforms to the requirements of item a) below, and either or both of items b) and c).

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

A claim of conformance shall identify the document that contains the description specified in a) below, and shall identify the adopted form(s), the adopted

iTeh STANDARD PREVIEW
(standards.iteh.ai)
ISO/IEC 10646-1:2000
https://standards.iteh.ai/catalog/standards/sist/23a1f63a-fe59-4266-aef0-8d25eaad12c2/iso-iec-10646-1-2000

implementation level, the adopted subset (by means of a list of collections and/or characters), and the selection of control functions adopted in accordance with clause 15.

a) Device description: A device that conforms to ISO/IEC 10646 shall be the subject of a description that identifies the means by which the user may supply characters to the device and/or may recognize them when they are made available to the user, as specified respectively, in 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 form and implementation level.

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

Any corresponding characters that are not within the adopted subset shall be indicated to the user. The way used for indicating them need not distinguish them from each other.

NOTE 1 - An indication to the user may consist of making available the same character to represent all characters not in the adopted subset, or providing a distinctive audible or visible signal when appropriate to the type of user.

NOTE 2 - See also annex J for receiving devices with retransmission capability.

3 Normative references

The following normative documents contain provisions which, through reference in this text, constitute provisions of this part 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 this part of 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.*

4 Terms and definitions

For the purposes of this part of ISO/IEC 10646, the following terms and definitions apply:

4.1 Basic Multilingual Plane (BMP): Plane 00 of Group 00.

4.2 block: A contiguous range of code positions to which a set of characters that share common characteristics, such as script, are allocated. A block does not overlap another block. One or more of the code positions within a block may have no character allocated to it.

4.3 canonical form: The form with which characters of this coded character set are specified using four octets to represent each character.

4.4 CC-data-element (coded-character-data-element): An element of interchanged information that is specified to consist of a sequence of coded representations of characters, in accordance with one or more identified standards for coded character sets.

iTeh STANDARD PREVIEW
(standards.iteh.ai)
ISO/IEC 10646-1:2000
https://standards.iteh.ai/catalog/standards/sist/23a1f63a-fe59-4266-aef0-8d25aaad12e2/iso-iec-10646-1-2000

4.5 cell: The place within a row at which an individual character may be allocated.

4.6 character: A member of a set of elements used for the organization, control, or representation of data.

4.7 character boundary: Within a stream of octets the demarcation between the last octet of the coded representation of a character and the first octet of that of the next coded character.

4.8 coded character: A character together with its coded representation.

4.9 coded character set: A set of unambiguous rules that establishes a character set and the relationship between the characters of the set and their coded representation.

4.10 code table: A table showing the characters allocated to the octets in a code.

4.11 collection: A set of coded characters which is numbered and named and which consists of those coded characters whose code positions lie within one or more identified ranges.

NOTE - If any of the identified ranges include code positions to which no character is allocated, the repertoire of the collection will change if an additional character is assigned to any of those positions 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.

4.12 combining character: A member of an identified subset of the coded character set of ISO/IEC 10646 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 4.14).

NOTE - This part of ISO/IEC 10646 specifies several subset collections which include combining characters.

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

4.14 composite sequence: A sequence of graphic characters consisting of a non-combining character followed by one or more combining characters (see also 4.12).

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 is not a character and therefore is not a member of the repertoire of ISO/IEC 10646.

4.15 control function: An action that affects the recording, processing, transmission, or interpretation of data, and that has a coded representation consisting of one or more octets.

4.16 default state: The state that is assumed when no state has been explicitly specified.

4.17 detailed code table: A code table showing the individual characters, and normally showing a partial row.

4.18 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.)

4.19 fixed collection: A collection in which every code position 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.

4.20 graphic character: A character, other than a control function, that has a visual representation normally handwritten, printed, or displayed.

4.21 graphic symbol: The visual representation of a graphic character or of a composite sequence.

4.22 group: A subdivision of the coding space of this coded character set; of 256 x 256 x 256 cells.

4.23 high-half zone: a set of cells reserved for use in UTF-16 (see annex C); an RC-element corresponding to any of these cells may be used in UTF-16 as the first of a pair of RC-elements which represents a character from a plane other than the BMP.

4.24 interchange: The transfer of character coded data from one user to another, using telecommunication means or interchangeable media.

4.25 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.

4.26 low-half zone: a set of cells reserved for use in UTF-16 (see annex C); an RC-element corresponding to any of these cells may be used in UTF-16 as the second of a pair of RC-elements which represents a character from a plane other than the BMP.

4.27 octet: An ordered sequence of eight bits considered as a unit.

4.28 plane: A subdivision of a group; of 256 x 256 cells

4.29 presentation; to present: The process of writing, printing, or displaying a graphic symbol.

4.30 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.

4.31 private use plane: A plane within this coded character set the contents of which is not specified in ISO/IEC 10646 (see clause 10)

4.32 RC-element: a two-octet sequence comprising the R-octet and the C-octet (see 6.2) from the four octet sequence (in the canonical form) that corresponds to a cell in the coding space of this coded character set.

4.33 repertoire: A specified set of characters that are represented in a coded character set.

4.34 row: A subdivision of a plane; of 256 cells.

4.35 script: A set of graphic characters used for the written form of one or more languages.

4.36 supplementary plane: A plane that accommodates characters which have not been allocated to the Basic Multilingual Plane.

4.37 unpaired RC-element: An RC-element in a CC-data element that is either:

- an RC-element from the high-half zone that is not immediately followed by an RC-element from the low-half zone, or
- an RC-element from the low-half zone that is not immediately preceded by a high-half RC-element from the high-half zone.

iTeh STANDARD PREVIEW
(standards.iteh.ai)
ISO/IEC 10646-1:2000
https://standards.iteh.ai/catalog/standards/sist/23a1f63a-fe59-4266-aef0-8d25aead12e2/iso-iec-10646-1-2000

4.38 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.)

4.39 zone: A sequence of cells of a code table, comprising one or more rows, either in whole or in part, containing characters of a particular class (for example see clause 8).

5 General structure of the UCS

The general structure of the Universal Multiple-Octet Coded Character Set (referred to hereafter as "this coded character set") is described in this explanatory clause, and is illustrated in figures 1 and 2. The normative specification of the structure is given in the following clauses.

The value of any octet is expressed in hexadecimal notation from 00 to FF in ISO/IEC 10646 (see annex K).

The canonical form of this coded character set – the way in which it is to be conceived – uses a four-dimensional coding space, regarded as a single entity, consisting of 128 three-dimensional groups.

NOTE - Thus, bit 8 of the most significant octet in the canonical form of a coded character can be used for internal processing purposes within a device as long as it is set to zero within a conforming CC-data-element.

Each group consists of 256 two-dimensional planes. Each plane consists of 256 one-dimensional rows, each row containing 256 cells. A character is located and coded at a cell within this coding space or the cell is declared unused.

In the canonical form, four octets are used to represent each character, and they specify the group, plane, row and cell, respectively. The canonical form consists of four octets since two octets are not sufficient to cover all the characters in the world, and a 32-bit representation follows modern processor architectures.

The four-octet canonical form can be used as a four-octet coded character set, in which case it is called UCS-4.

The first plane (Plane 00 of Group 00) is called the Basic Multilingual Plane. The Basic Multilingual Plane includes characters in general use in alphabetic, syllabic, and ideographic scripts together with various symbols and digits.

The subsequent planes are regarded as supplementary or private use planes, which will accommodate additional graphic characters (see clause 9).

The planes that are reserved for private use are specified in clause 10. The contents of the cells in private use zones are not specified in ISO/IEC 10646.

Each character is located within the coded character set in terms of its Group-octet, Plane-octet, Row-octet, and Cell-octet.

In addition to the canonical form, a two-octet BMP form is specified. Thus, the Basic Multilingual Plane can be used as a two-octet coded character set identified as UCS-2.

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

A UCS Transformation Format (UTF-16) is specified in annex C which can be used to represent characters from 16 planes of group 00, additional to the BMP, in a form that is compatible with the two-octet BMP form.

Another UCS Transformation Format (UTF-8) is specified in annex D which can be used to transmit text data through communication systems which are sensitive to octet values for control characters coded according to the 8-bit structure of ISO/IEC 2022, and to ISO/IEC 4873. UTF-8 also avoids the use of octet values according to ISO/IEC 4873 which have special significance during the parsing of file-name character strings in widely-used file-handling systems.

iTeh STANDARD PREVIEW
(standards.iteh.ai)
ISO/IEC 10646-1:2000
https://standards.iteh.ai/catalog/standards/sist/23a1f63a-fe59-4266-ae10-8d25cad1d2e2/iso-iec-10646-1-2000

6 Basic structure and nomenclature

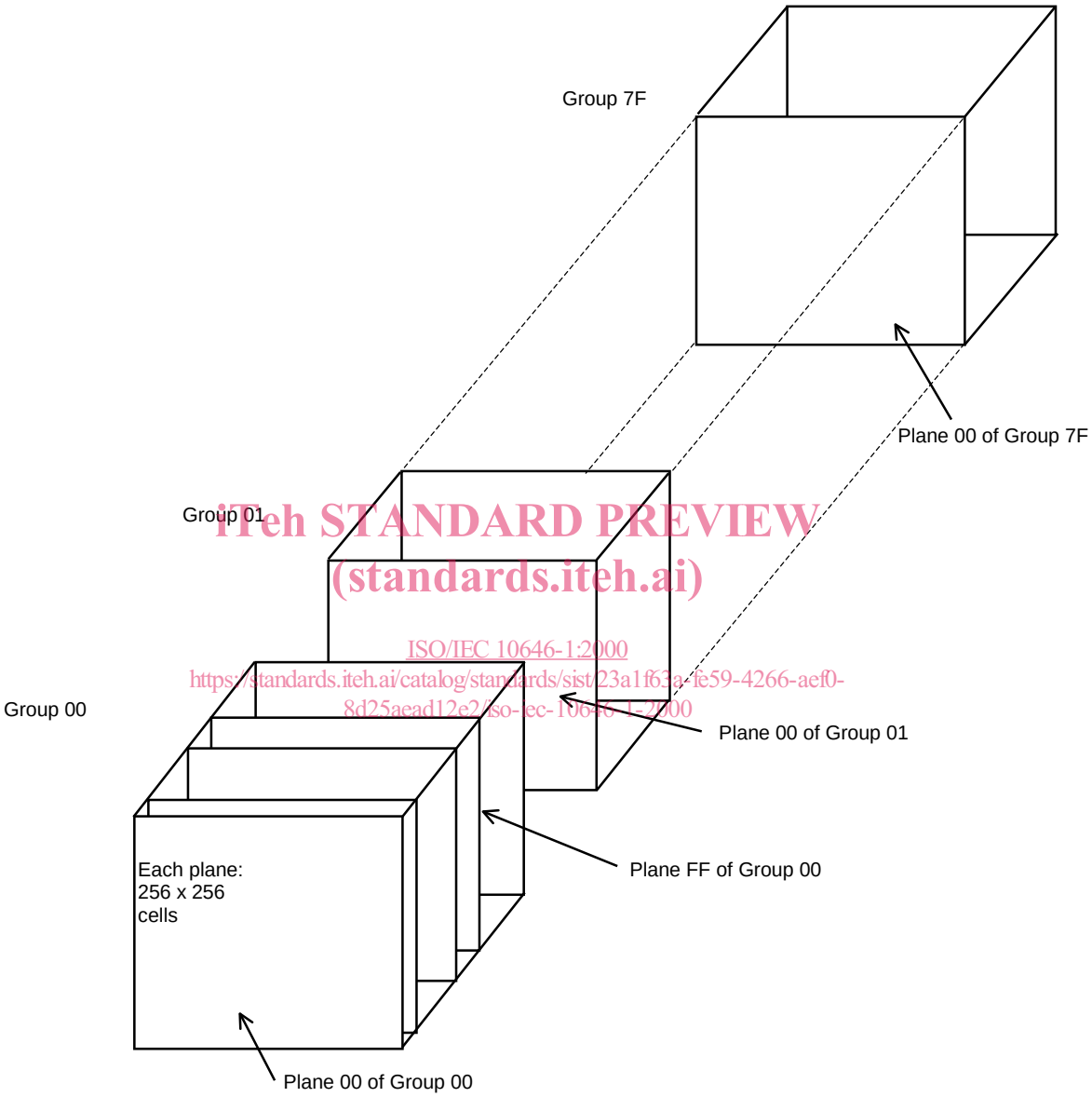
6.1 Structure

The Universal Multiple-Octet Coded Character Set as specified in ISO/IEC 10646 shall be regarded as a single entity.

This entire coded character set shall be conceived of as comprising 128 groups of 256 planes. Each plane shall be regarded as containing 256 rows of characters, each row containing 256 cells. In a code table representing the contents of a plane (such as in figure 2), the horizontal axis shall represent the least significant octet, with its smaller value to the left; and the vertical axis shall represent the more significant octet, with its smaller value at the top.

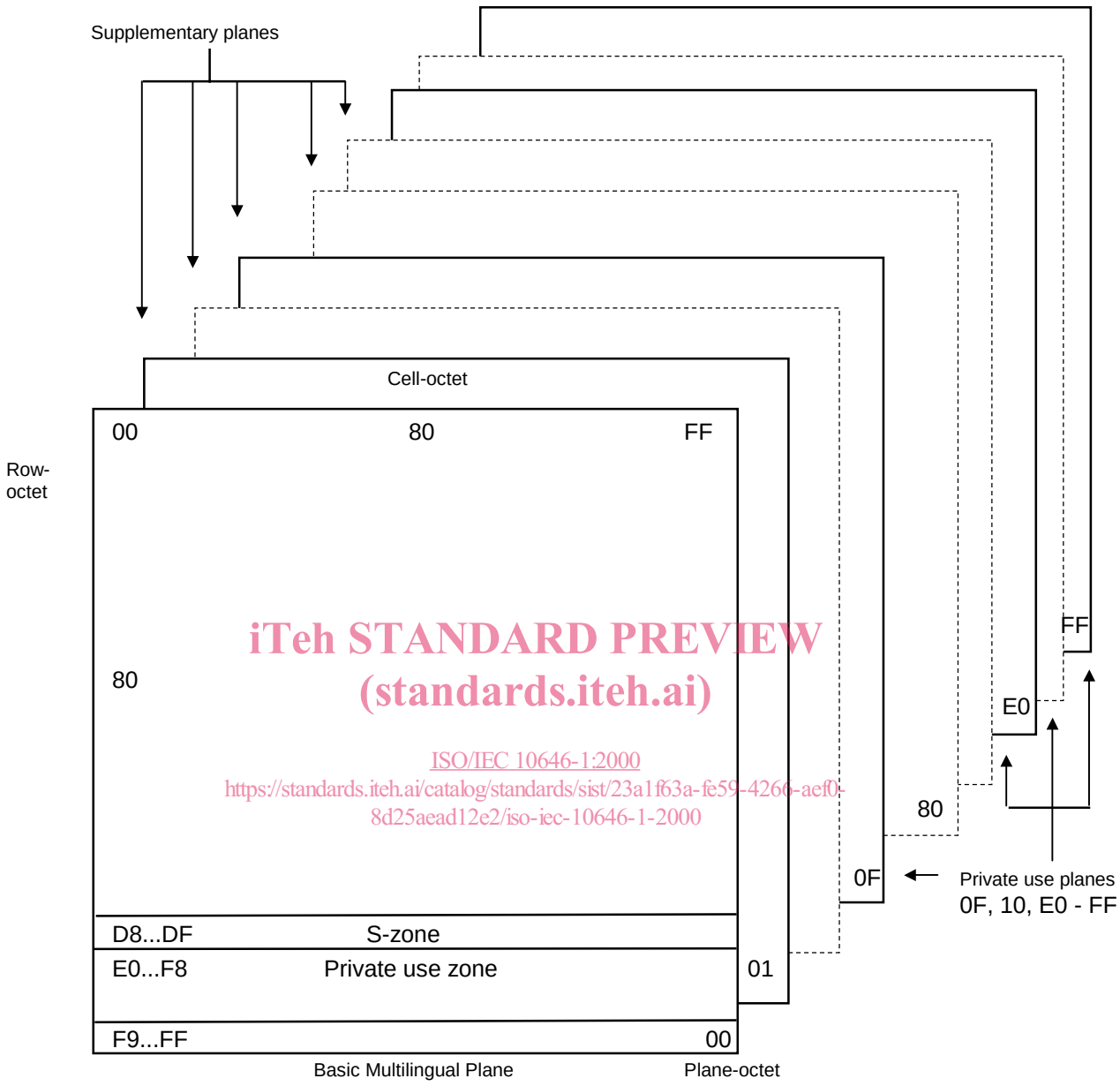
Each axis of the coding space shall be coded by one octet. Within each octet the most significant bit shall be bit 8 and the least significant bit shall be bit 1. Accordingly, the weight allocated to each bit shall be

bit 8	bit 7	bit 6	bit 5	bit 4	bit 3	bit 2	bit 1
128	64	32	16	8	4	2	1



iTeh STANDARD PREVIEW
(standards.iteh.ai)
ISO/IEC 10646-1:2000
https://standards.iteh.ai/catalog/standards/sist/23a1f63a-fe59-4266-aef0-8d25aead12e2/iso-iec-10646-1-2000

Figure 1 - Entire coding space of the Universal Multiple-Octet Coded Character Set



iTeh STANDARD PREVIEW
(standards.iteh.ai)
ISO/IEC 10646-1:2000
https://standards.iteh.ai/catalog/standards/sist/23a1f63a-fe59-4266-aef0-8d25aead12e2/iso-iec-10646-1-2000

NOTE - Labels “S-zone” and “Private use zone” are specified in clause 8.

Figure 2 - Group 00 of the Universal Multiple-Octet Coded Character Set

6.2 Coding of characters

In the canonical form of the coded character set, each character within the entire coded character set shall be represented by a sequence of four octets. The most significant octet of this sequence shall be the group-octet. The least significant octet of this sequence shall be the cell-octet. Thus this sequence may be represented as

m.s. l.s.

Group-octet	Plane-octet	Row-octet	Cell-octet

where m.s. means the most significant octet, and l.s. means the least significant octet.

For brevity, the octets may be termed

m.s. l.s.

G-octet	P-octet	R-octet	C-octet

Where appropriate, these may be further abbreviated to G, P, R, and C.

The value of any octet shall be represented by two hexadecimal digits, for example: 31 or FE. When a single character is to be identified in terms of the values of its group, plane, row, and cell, this shall be represented such as:

0000 0030 for DIGIT ZERO

0000 0041 for LATIN CAPITAL LETTER A

When referring to characters within an identified plane, the leading four digits (for G-octet and P-octet) may be omitted. For example, within plane 00, 0030 may be used to refer to DIGIT ZERO.

6.3 Octet order

The sequence of the octets that represent a character, and the most significant and least significant ends of it, shall be maintained as shown above. When serialized as octets, a more significant octet shall precede less significant octets. When not serialized as octets, the order of octets may be specified by agreement between sender and recipient (see 16.1 and annex H).

6.4 Naming of characters

ISO/IEC 10646 assigns a unique name to each 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 clause 27 for Chinese /Japanese/Korean (CJK) unified ideographs.

Guidelines to be used for constructing the names of characters in cases a. and b. are given in annex L.

6.5 Short identifiers for characters

ISO/IEC 10646 defines a short identifier for each character. The short identifier for any character is distinct from the short identifier for any other character.

NOTE - 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 eight-digit form of short identifier shall consist of the sequence of eight hexadecimal digits that represents the code position of the character (see 6.2).

b. The four-digit form of short identifier shall consist of the last four digits of the eight-digit form. It is not defined if the first four digits of the eight-digit form are not all zeroes; that is, for characters allocated outside the Basic Multilingual Plane.

c. The character "-" (HYPHEN-MINUS) may, as an option, precede the 8-digit form of short identifier.

d. The character "+" (PLUS SIGN) may, as an option, precede the 4-digit form of short identifier.

e. The prefix letter "U" (LATIN CAPITAL LETTER U) may, as an option, precede any of the four forms of short identifier defined in a. to d. 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 | {-}xxxxxxxx]

where "x" represents one hexadecimal digit (0 to 9, A to F, or a to f), for example:

-hhhhhhhh +kkkk
Uhhhhhhhh U+kkkk

where hhhhhhhh indicates the eight-digit form and kkkk indicates the four-digit form.

NOTE 1 - As an example the short identifier for LATIN SMALL LETTER LONG S (see tables for Row 01 in clause 26) may be notated in any of the following forms:

0000017F -0000017F U0000017F U-0000017F
017F +017F U017F U+017F

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

NOTE 2 - Two special prefixed forms of notation have also been used, in which the letter T (LATIN CAPITAL LETTER T or LATIN SMALL LETTER T) replaces the

iTeh STANDARD PREVIEW
(standards.iteh.ai)
ISO/IEC 10646-1:2000
https://standards.iteh.ai/catalog/standards/sist/23a1f63a-fe59-4266-aef0-8d25acad12e2/iso-iec-10646-1-2000

letter U in the corresponding prefixed forms. The forms of notation that included the prefix letter T indicated that the short identifier refers to a character in ISO/IEC 10646-1 First Edition (before the application of any Amendments), whereas the forms of notation that include the prefix letter U always indicate that the short identifier refers to a character in ISO/IEC 10646 at the most recent state of amendment. Corresponding short identifiers of the form T-xxxxxxxx and U-xxxxxxxx refer to the same character except when xxxxxxxx lies in the range 00003400 to 00004DFF inclusive. Forms of notation that include no prefix letter always indicate a reference to the most recent state of amendment of ISO/IEC 10646, unless otherwise qualified.

7 General requirements for the UCS

The following requirements apply to the entire coded character set.

a) The values of P-, and R-, and C-octets used for representing graphic characters shall be in the range 00 to FF. The values of G-octets used for representation of graphic characters shall be in the range 00 to 7F. On any plane, code positions FFFE and FFFF shall not be used.

NOTE - Code position FFFE is reserved for "signature" (see annex H). Code position FFFF can be used for internal processing uses requiring a numeric value that is guaranteed not to be a coded character, such as in terminating tables, or signaling end-of-text. Since it is the largest two-octet value, it may also be used as the final value in binary or sequential searching index.

b) Code positions to which a character is not allocated, except for the positions reserved for private use characters or for transformation formats, 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 positions reserved for private use characters or for transformation formats.

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

8 The Basic Multilingual Plane

Plane 00 of Group 00 shall be the Basic Multilingual Plane (BMP). The BMP can be used as a two-octet coded character set in which case it shall be called UCS-2 (see 13.1).

Code positions 0000 0000 to 0000 001F in the BMP are reserved for control characters, and code position 0000 007F is reserved for the character DELETE (see clause 15). Code positions 0000 0080 to 0000 009F are reserved for control characters.

Code positions 0000 D800 to 0000 DFFF are reserved for the use of UTF-16 (see annex C). These positions are known as the S-zone.

Code positions 0000 E000 to 0000 F8FF are reserved for private use (see clause 10). These positions are known as the private use zone.

Code postions 0000 FFFE and 0000 FFFF are reserved.

9 Other planes

9.1 Planes reserved for future standardization

Planes 11 to DF in Group 00 and Planes 00 to FF in Groups 01 to 5F are reserved for future standardization, and thus those code positions shall not be used for any other purpose.

9.2 Planes accessible by UTF-16

Each code position in Planes 01 to 10 of Group 00 has a unique mapping to a four-octet sequence in accordance with the UTF-16 form of coded representation (see annex C). This form is compatible with the two-octet BMP form of UCS-2 (see 13.1).

Code positions in Planes 11 to FF of Group 00, or in Planes 00 to FF of other groups, do not have a mapping to the UTF-16 form.

iTeh STANDARD PREVIEW
(standards.iteh.ai)
ISO/IEC 10646-1:2000
https://standards.iteh.ai/catalog/standards/sist/23a1f63a-fe59-4266-aef0-ff25dea812e2/iso-iec-10646-1-2000

10 Private use groups, planes, and zones

10.1 Private use characters

Private use characters are not restrained 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 1 - For meaningful interchange of private use characters, an agreement, independent of ISO/IEC 10646, is necessary between sender and recipient.

Private use characters can be used for dynamically-redefinable character applications.

NOTE 2 - For meaningful interchange of dynamically-redefinable characters, an agreement, independent of ISO/IEC 10646 is necessary between sender and recipient. ISO/IEC 10646 does not specify the techniques for defining or setting up dynamically-redefinable characters.

10.2 Code positions for private use characters

The code positions of the 32 groups from Group 60 to Group 7F shall be for private use.

The code positions of Plane 0F and Plane 10, and of the 32 planes from Plane E0 to Plane FF, of Group 00 shall be for private use.

The 6400 code positions E000 to F8FF of the Basic Multilingual Plane shall be for private use.

The contents of these code positions are not specified in ISO/IEC 10646 (see 10.1).

11 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.

12 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.

12.1 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 interwork with this coded character set.

A claim of conformance referring to a limited subset shall list the graphic characters in the subset by the names of graphic characters or code positions as defined in ISO/IEC 10646.

12.2 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 an annex of each part of ISO/IEC 10646 (see annex A). A selected subset shall always automatically include the Cells 20 to 7E of Row 00 of Plane 00 of Group 00.

A claim of conformance referring to a selected subset shall list the collections chosen as defined in ISO/IEC 10646.

13 Coded representation forms of the UCS

ISO/IEC 10646 provides four alternative forms of coded representation of characters. Two of these forms are specified in this clause, and two others, UTF-16 and UTF-8, are specified in annexes C and D respectively.

NOTE - The characters from the ISO/IEC 646 IRV repertoire are coded by simple zero extensions to their coded representations in ISO/IEC 646 IRV. Therefore, their coded representations have the same integer values when represented as 8-bit, 16-bit, or 32-bit integers. For implementations sensitive to a zero-valued octet (e.g. for use as a string terminator), use of 8-bit based array data type should be avoided as any zero-valued octet may be interpreted incorrectly. Use of data types at least 16-bits wide is more suitable for UCS-2, and use of data types at least 32-bits wide is more suitable for UCS-4.

13.1 Two-octet BMP form

This coded representation form permits the use of characters from the Basic Multilingual Plane with each character represented by two octets.

Within a CC-data-element conforming to the two-octet BMP form, a character from the Basic Multilingual Plane shall be represented by two octets comprising the R-octet and the C-octet as specified in 6.2 (i.e. its RC-element).

NOTE - A coded graphic character using the two-octet BMP form may be implemented by a 16-bit integer for processing.

13.2 Four-octet canonical form

The canonical form permits the use of all the characters of ISO/IEC 10646, with each character represented by four octets.

Within a CC-data-element conforming to the four-octet canonical form, every character shall be represented by four octets comprising the G-octet, the P-octet, the R-octet, and the C-octet as specified in 6.2.

NOTE - A coded graphic character using the four-octet canonical form may be implemented by a 32-bit integer for processing.

iTeh STANDARD PREVIEW
(standards.iteh.ai)
ISO/IEC 10646-1:2000
https://standards.iteh.ai/catalog/standards/sist/23a1f63a-fe59-4266-acf0-8d25aead12e2/iso-iec-10646-1-2000

14 Implementation levels

ISO/IEC 10646 specifies three levels of implementation. Combining characters are described in 24 and listed in annex B.

14.1 Implementation level 1

When implementation level 1 is used, a CC-data-element shall not contain coded representations of combining characters (see clause B.1) nor of characters from HANGUL JAMO block (see clause 25). When implementation level 1 is used the unique-spelling rule shall apply (25.2).

14.2 Implementation level 2

When implementation level 2 is used, a CC-data-element shall not contain coded representations of characters listed in clause B.2. When implementation level 2 is used the unique-spelling rule shall apply (25.2).