SLOVENSKI STANDARD
SIST ISO/IEC 10646:2008

01-november-2008

Informacijska tehnologija ï Univerzalni veļ oktetni nabor znakov (UCS)

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

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

Ta slovenski standard je istoveten z: **ISO/IEC 10646:2003**

ICS:

35.040	Nabori znakov in kodiranje informacij	Character sets and information coding

SIST ISO/IEC 10646:2008 **en,fr**

INTERNATIONAL STANDARD

ISO/IEC 10646

First edition
2003-12-15

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

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

Reference number
ISO/IEC 10646:2003(E)

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

Published in Switzerland

Contents

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

Annexes

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

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

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

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

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

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

Introduction

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

By defining a consistent way of encoding multilingual text it enables the exchange of data internationally. The information technology industry gains data stability, greater global interoperability and data interchange. ISO/IEC 10646 has been widely adopted in new Internet protocols and implemented in modern operating systems and computer languages. This edition covers over 95 000 characters from the world's scripts.

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

— CJKU_SR.txt

— CJKC_SR.txt

— Allnames.txt

— HangulX.txt

— HangulSy.txt

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

1 Scope

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

This document:

- specifies the architecture of ISO/IEC 10646,
- defines terms used in ISO/IEC 10646,
- describes the general structure of the coded character set;
- specifies the Basic Multilingual Plane (BMP) of the UCS,
- specifies supplementary planes of the UCS: the Supplementary Multilingual Plane (SMP), the Supplementary Ideographic Plane (SIP) and the Supplementary Special-purpose Plane (SSP),
- defines a set of graphic characters used in scripts and the written form of languages on a world-wide scale;
- specifies the names for the graphic characters of the BMP, SMP, SIP, SSP and their coded representations;
- specifies the four-octet (32-bit) canonical form of the UCS: UCS-4;
- specifies a two-octet (16-bit) BMP form of the UCS: UCS-2;
- specifies the coded representations for control functions;
- specifies the management of future additions to this coded character set.

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

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

NOTE – The Unicode Standard, Version 4.0 includes a set of characters, names, and coded representations that are identical with those in this International Standard. It additionally provides details of character properties, processing algorithms, and definitions that are useful to implementers.

2 Conformance

2.1 General

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

2.2 Conformance of information interchange

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

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

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

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

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

2.3 Conformance of devices

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

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

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

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

b) **Originating device**: An originating device shall allow its user to supply any characters from an adopted subset, and be capable of transmitting their coded representations within a CC-data-element in accordance with the adopted form and implementation level.

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

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

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

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

3 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

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

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

Unicode Standard Annex, UAX#9, The Unicode Bidirectional Algorithm, Version 4.0.0, 2003-04-17.

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

4 Terms and definitions

For the purposes of this document, 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 a 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 them.

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.

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 – 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 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.

4.27 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.

4.28 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.

4.29 Octet

An ordered sequence of eight bits considered as a unit.

4.30 Plane

A subdivision of a group; of 256 x 256 cells.

4.31 Presentation; to present

The process of writing, printing, or displaying a graphic symbol.

4.32 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.

4.33 Private use plane

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

4.34 RC-element
A two-octet sequence comprising the R-octet and the C-octet (see clause 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.

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

4.36 Row
A subdivision of a plane; of 256 cells.

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

4.38 Supplementary plane
A plane other than Plane 00 of Group 00; a plane that accommodates characters which have not been allocated to the Basic Multilingual Plane.

4.39 Supplementary Multilingual Plane for scripts and symbols (SMP)
Plane 01 of Group 00.

4.40 Supplementary Ideographic Plane (SIP)
Plane 02 of Group 00.

4.41 Supplementary Special-purpose Plane (SSP)
Plane 0E of Group 00.

4.42 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 an RC-element from the high-half zone.

4.43 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.)

4.44 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 1 – 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.

NOTE 2 – The use of the term “canonical” for this form does not imply any restriction or preference for this form over transformation formats that a conforming implementation may choose for the representation of UCS characters.

ISO/IEC 10646 defines graphic characters and their coded representation for the following planes:

- The Basic Multilingual Plane (BMP, Plane 00 of Group 00). The Basic Multilingual Plane can be used as a two-octet coded character set identified as UCS-2.
- The Supplementary Multilingual Plane for scripts and symbols (SMP, Plane 01 of Group 00).
- The Supplementary Ideographic Plane (SIP, Plane 02 of Group 00).
- The Supplementary Special-purpose Plane (SSP, Plane 0E of Group 00).

Additional supplementary planes may be defined in the future to accommodate additional graphic characters.

The planes that are reserved for private use are specified in clause 10. The contents of the cells in private use planes and 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.

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 supplementary planes of Group 00 (Planes 01 to 10), in addition to the BMP (Plane 00), 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 that have special significance during the parsing of file-name character strings in widely-used file-handling systems.

6 Basic structure and nomenclature

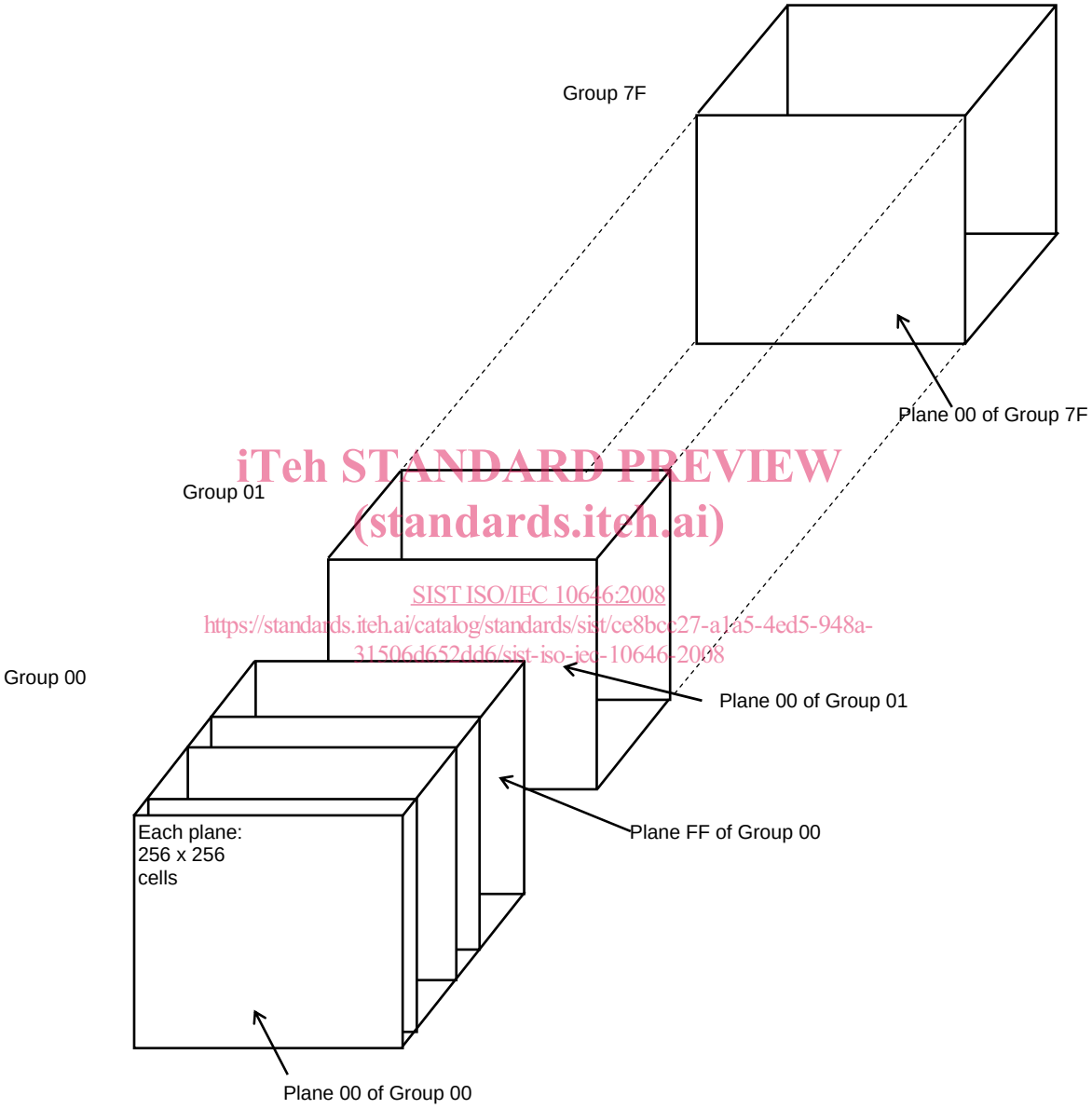
6.1 Structure

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.

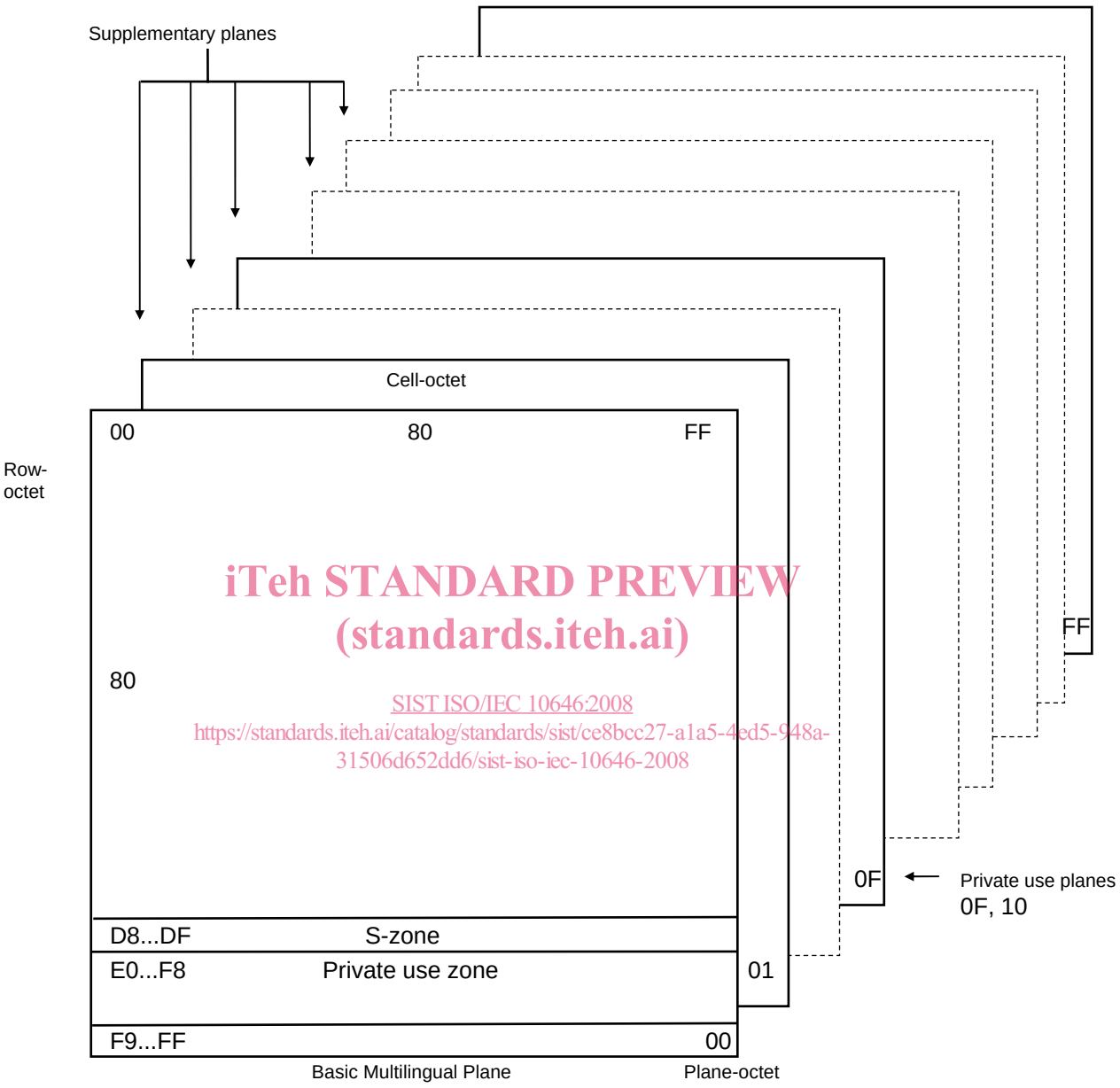
Each axis of the coding space shall be coded by one octet. Within each octet the most significant bit shall be bit 8 and the least significant bit shall be bit 1. Accordingly, the weight allocated to each bit shall be:

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



NOTE – To ensure continued interoperability between the UTF-16 form and other coded representations of the UCS, it is intended that no characters will be allocated to code positions in Planes 11 to FF in Group 00 or any planes in any other groups.

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



NOTE 1 – Labels “S-zone” and “Private use zone” are specified in clause 8.

NOTE 2 – To ensure continued interoperability between the UTF-16 form and other coded representations of the UCS, it is intended that no characters will be allocated to code positions in Planes 11 to FF in Group 00.

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