INTERNATIONAL STANDARD

ISO/IEC 10646-1

First edition
1993-05-01

AMENDMENT 1
1996-10-15

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

Part 1:
Architecture and Basic Multilingual Plane

AMENDMENT 1: Transformation Format for 16 planes of group 00 (UTF-16)

Technologies de l'information — Jeu universel de caractères codés à plusieurs octets —

Partie 1: Architecture et table multilingue

AMENDEMENT 1: Format de transformation pour 16 tables du groupe 00 (UTF-16)

Reference number
ISO/IEC 10646-1:1993/Amd.1:1996(E)

Contents

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. 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.

Amendment 1 to International Standard ISO/IEC 10646-1:1993 was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*.

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 amendment to ISO/IEC 10646 specifies an additional transformation format, UTF-16. UTF-16 is a coded representation that permits over a million graphic characters of the UCS to be represented in a form which is compatible with the two-octet BMP form.

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

Part 1:

Architecture and Basic Multilingual Plane

AMENDMENT 1: Transformation Format for 16 planes of group 00 (UTF-16)

2 Conformance

Clause 2 applies with the text of 2.2 a) amended to read:

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 14 or Annex Q, and to an identified implementation level chosen from clause 15;

4 Definitions

Clause 4 applies with the following additions and amendments:

Renumber 4.21 - 4.22 as 4.22 - 4.23

Renumber 4.23 - 4.27 as 4.25 - 4.29

Renumber 4.28 - 4.31 as 4.31 - 4.34

Renumber 4.32 - 4.33 as 4.36 - 4.37

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

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

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

4.35 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.

5 General structure of the UCS

Clause 5 applies with the text amended as follows. In the sixth paragraph on page 4, replace:

The 32 planes with Plane-octet values E0 to FF of Group 00 are for Private Use. The 32 groups with Group-octet values 60 to 7F of this coded character set are also for Private Use.

with:

The planes that are reserved for Private Use are specified in clause 11.

and add the following new paragraph at the end of the clause:

A UCS Transformation Format (UTF-16) is specified in Annex Q 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.

7 Special features of the UCS

Clause 7 applies with the text in paragraph 2 amended to read:

2. 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 standardisation 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.

8 The Basic Multilingual Plane

Clause 8 applies with the text amended as follows, and the diagram replaced with an amended version as shown. Replace:

The Basic Multilingual Plane shall be divided into four zones:

A-zone: code positions 0000 to 4DFF

I-zone: code positions 4E00 to 9FFF

O-zone: code positions A000 to DFFF

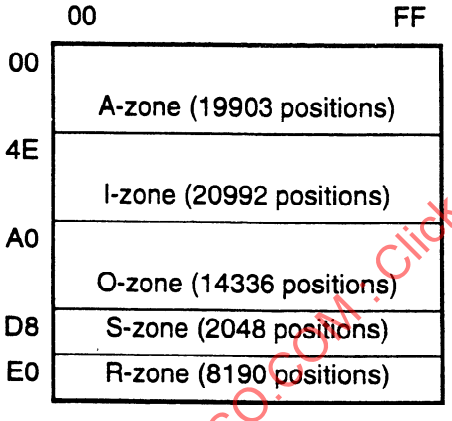
R-zone: code positions E000 to FFFD

with:

The Basic Multilingual Plane shall be divided into five zones:

zone	code positions
A-zone:	0000 0000 to 0000 4DFF
I-zone:	0000 4E00 to 0000 9FFF
O-zone:	0000 A000 to 0000 D7FF
S-zone:	0000 D800 to 0000 DFFF
R-zone:	0000 E000 to 0000 FFFD

The amended version of the diagram is as follows:



Replace:

Code positions 0000 to 001F in the BMP are reserved for control characters, and code position 007F is reserved for the character DELETE (see clause 16). Code positions 0080 to 009F are reserved.

with:

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 16). Code positions 0000 0080 to 0000 009F are reserved.

In the last paragaph, after:

The O-zone is reserved for future standardisation.

insert:

The S-zone is reserved for the use of UTF-16 (see Annex Q).

9 Other planes

Clause 9 is amended to read:

9.1 Planes reserved for future standardisation

Planes 11 to DF in Group 00 and planes 00 to FF in Groups 01 to 5F are reserved for future standardisation, 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 Q). This form is compatible with the two-octet BMP form of UCS-2 (see 14.1).

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.

11 Private Use groups and planes

Clause 11 applies with the text amended as follows. Replace:

The code positions of 32 planes from Plane E0 to Plane FF of Group 00 shall be for Private Use.

with:

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.

14.1 Two-octet BMP form

Clause 14.1 applies with the text amended as follows. At the end of the second paragraph replace:

in 6.2.

with:

in 6.2 (i.e. its RC-element).

Annex A - Collections of graphic characters for subsets

Annex A applies with the text amended as follows.
Replace:

300 BMP 0000 - FFFD

with:
300 BMP 0000 - D7FF
E000 - FFFD

Replace:

400 PRIVATE USE PLANES G=00, P=E0-FF

with:
400 PRIVATE USE PLANES G=00,
P=0F, 10 & E0-FF

Annex F - The use of "signatures" to identify UCS

Annex F applies with the text amended as follows.
After:

UCS-4 signature: 0000 FEFF

insert:
UTF-16 signature: FEFF

Annex M - External references to character repertoires

Annex M applies with M.3 amended as follows. In the third paragraph replace:

- four-octet-form (4).

with:

- four-octet-form (4), or
- UTF16 (5).

Replace:

"ISO 10646 part-1 form 2" and "ISO 10646 part-1 form 4".

with:

- "ISO 10646 part-1 form 2"
- "ISO 10646 part-1 form 4"
- "ISO 10646 part-1 utf-16".

Add the following new annex:

Annex Q

(normative)

Transformation format for 16 planes of Group 00 (UTF-16)

UTF-16 provides a coded representation of over a million graphic characters of UCS-4 in a form that is compatible with the two-octet BMP form of UCS-2 (14.1). This permits the coexistence of those characters from UCS-4 within coded character data that is in accordance with UCS-2.

In UTF-16 each graphic character from the UCS-2 repertoire retains its UCS-2 coded representation. In addition, the coded representation of any character from a single contiguous block of 16 Planes in Group 00 (1,048,576 code positions) consists of a pair of RC-elements (4.30), where each such RC-element corresponds to a cell in a single contiguous block of 8 Rows in the BMP (2,048 code positions). These code positions are reserved for the use of this coded representation form, and shall not be allocated for any other purpose.

Q.1 Specification of UTF-16

The specification of UTF-16 is as follows:

1. The high-half zone shall be the 4 rows D8 to DB of the BMP, i.e., the 1,024 cells in the S-zone whose code positions are from D800 through DBFF.
2. The low-half zone shall be the 4 rows DC to DF of the BMP, i.e., the 1,024 cells in the S-zone whose code positions are from DC00 through DFFF.
3. All cells in the high-half zone and the low-half zone shall be permanently reserved for the use of the UTF-16 coded representation form.
4. In UTF-16, any UCS character from the BMP shall be represented by its UCS-2 coded representation as specified by the body of this international standard.
5. In UTF-16, any UCS character whose UCS-4 coded representation is in the range 0001 0000 to 0010 FFFF shall be represented by a sequence of two RC-elements from the S-zone, of which the first is an RC-element from the high-half zone, and the second is an RC-element from the low-half zone.

   The mapping between UCS-4 and UTF-16 for these characters shall be as shown in Q.3; the reverse mapping is shown in Q.4.

Q.2 Notation

1. All numbers are in hexadecimal notation.
2. Double-octet boundaries in the notations for UTF-16 are indicated with semicolons.
3. The symbol "%" indicates the modulo operation, e.g.: x % y = x modulo y.
4. The symbol "/" indicates the integer division operation, e.g.: 7 / 3 = 2.
5. Precedence is integer-division > modulo-operation > integer-multiplication > integer-addition.

Q.3 Mapping from UCS-4 form to UTF-16 form

UCS-4 (4-octet)		UTF-16, 2-octet elements
x =	0000 0000 .. 0000 FFFF (see Note 1)	x % 0001 0000;
x =	0001 0000 .. 0010 FFFF	y; z;
where	y = ((x - 0001 0000) / 400) + D800	
	z = ((x - 0001 0000) % 400) + DC00	
x	0011 0000 .. 7FFF FFFF	(no mapping (is defined

NOTE 1 - Code positions from 0000 D800 to 0000 DFFF are reserved for the UTF-16 form and do not occur in UCS-4. The values 0000 FFFE and 0000 FFFF also do not occur (see clause 8). The mapping of these code positions in UTF-16 is undefined.

Example:

The UCS-4 sequence [0000 0048] [0000 0069] [0001 0000] [0000 0021] [0000 0021]

represents "Hi<0001 0000>!!".

It is mapped to UTF-16 as:

[0048] [0069] [D800] [DC00] [0021] [0021]

If interpreted as UCS-2 this sequence will be

"Hi<RC-element from high-half zone>
<RC-element from low-half zone>!!"

Q.4 Mapping from UTF-16 form to UCS-4 form

UTF-16, 2-octet elements		UCS-4 (4-octet)
x =	0000; .. D7FF;	x
x =	E000; .. FFFF;	x
pair (x, y) such that		
x =	D800; .. DBFF;	((x - D800) * 400
y =	DC00; .. DFFF;	+ (y - DC00))
		+ 0001 0000

Example:

The UTF-16 sequence

[0048] [0069] [D800] [DC00] [0021] [0021]

is mapped to UCS-4 as

[0000 0048] [0000 0069] [0001 0000]
[0000 0021] [0000 0021]

and represents "Hi<0001 0000>!!".

Q.5 Identification of UTF-16

When the escape sequences from ISO/IEC 2022 are used, the identification of UTF-16 and an implementation level (see clause 15) shall be by a designation sequence chosen from the following list:

ESC 02/05 02/15 04/10

UTF-16 with implementation level 1

ESC 02/05 02/15 04/11

UTF-16 with implementation level 2

ESC 02/05 02/15 04/12

UTF-16 with implementation level 3

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 16.

When the escape sequences from ISO/IEC 2022 are used, the identification of a return, or transfer, from UTF-16 to the coding system of ISO/IEC 2022 shall be as specified in 17.5 for a return or transfer from UCS.

Q.6 Unpaired RC-elements: Interpretation by receiving devices

According to Q.1 an unpaired RC-element (4.35) is not in conformance with the requirements of UTF-16.

If a receiving device that has adopted the UTF-16 form receives an unpaired RC-element because of error conditions either:

- in an originating device, or
- in the interchange between an originating and the receiving device, or
- in the receiving device itself,

then it shall interpret that unpaired RC-element in the same way that it interprets a character that is outside the adopted subset that has been identified for the device (see 2.3c).

NOTE 2 - Since a high-half RC-element followed by a low-half RC-element is a sequence that is in accordance with UTF-16, the only possible type of syntactically malformed sequence is an unpaired RC-element.

Example:

A receiving/originating device which only handles the Latin-1 repertoire, and uses boxes to display missing glyphs would display:

"The Greek letter <alpha> corresponds
to<hieroglyphicHigh>."

as:

"The Greek letter <box> corresponds to<box>."

Accordingly a similar device that can also interpret a UTF-16 data stream should display an unpaired RC-element as a <box> also.

Q.7 Receiving devices, advisory notes

When a receiving device interprets a CC-data-element that is in accordance with UTF-16 the following advisory notes apply.

1. UTF-16 is designed to be compatible with the UCS-2 two-octet BMP Form (14.1). The high-half and low-half zones are assigned to separate ranges of code positions, to which characters can never be assigned. Thus the function of every RC-element (two-octet unit) within a UTF-16 data stream is always immediately identifiable from its value, without regard to context.

   For example, the valid UTF-16 sequence [0048] [0069] [D800] [DC00] [0021] [0021] may also be interpreted, by a receiving device, that has adopted only UCS-2, as the coded representation of

   "Hi<unrecognized><unrecognized>!!"

   This form of compatibility is possible because RC-elements from the S-zone are interpreted