

ISO/IEC JTC1/SC2/WG2 N

Title: Missing Alternate Format Characters in ISO/IEC 10646

Author: V.S. Umamaheswaran (umavs@ca.ibm.com)

Date: 2003-10-13

Comments received during national body review of the working draft of the joint edition of 10646 revealed that the list of Alternate Format Characters was missing some characters. This document is the result of an investigation into missing entries from the list.

While trying to come up with the list, I realized that there is no definition of what constitutes an Alternate Format Character. Neither 10646, nor Unicode have a formal definition of the term. Unicode describes characters with 'unusual properties' in section 4.11.

The opening paragraph of Annex F – Alternate Format Characters, in 10646 (joint edition) is as follows:

“There is a special class of characters called Alternate Format Characters which are included for compatibility with some industry practices. These characters do not have printable graphic symbols, and are thus represented in the character code tables by dotted boxes.”

To be able to find out which characters qualify to be declared as Alternate Format Characters (AFC), I started with the clue that their glyphs will have a dotted box in the charts. Not all glyphs with dotted boxes are Alternate Format Characters either. All the characters with the general category of Cf did have a dotted box for their glyphs. There were several others with the dotted box with general category value other than Cf.

I am not sure if there are characters that may qualify for being Alternate Format Characters having a glyph other than with a dotted / dashed square box in them. If there are any, it looks like there is no easy way of extracting these from the standard.

Some observations:

- a. All Control Characters – having the general category Cc – are not AFCs.
- b. Space characters are listed separately. Some characters with Zs property are in the AFC list.
- c. Code position ranges reserved for AFCs have been defined in 10646, including some collection numbers. Variation Selection Characters -1 to -256 have been called out as Alternate Format Characters. Script-specific VS-s are not called AFCs.
- d. Several Cf category characters are not called or listed as AFCs.

Relevant texts from Joint Edition:

- From Clause 8 –
Code positions 0000 2060 to 0000 206F, 0000 FFF0 to 0000 FFFC, and 000E 0000 to 000E 0FFF are reserved for Alternate Format Characters (see annex F).
NOTE 2 – Unassigned code positions in those ranges may be ignored in normal processing and display.
- Clause 20.1 is a list of Space Characters
- Clause 20.3 is a list of Alternate Format Characters and are (supposed to be) described further in Annex F.
- From Clause 32 on SSP:
Code positions from E0000 to E0FFF are reserved for Alternate Format Characters (see clause 20).
- From Annex A:

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
3002	ALTERNATE FORMAT CHARACTERS	E0000-E0FFF

Candidates for AFC additions

The following table contains all characters having glyphs in the charts with a dashed / dotted box in them. The table contains in col 1 - the code position; in col 2 - the Page # from the hardcopy of Version 4 of Unicode book along with its block name; in col 3 - individual character names or group name to indicate what the list of code positions represent; in col 4 – the general category property from the Unicode database; in col 5 – clause number etc. where some mention of that code position / group of code positions appear in the joint edition text of 10646. The last column, col 6, indicates by ‘???’ entry that the character is a candidate for adding as an AFC in 10646.

Code Positions	p# from V4 book; Block name	Character name	Gen. Cat. ****	From Joint Edition Text	Candidate for AFC list / annex F
0000 – 001F	p420; C0 Controls and Basic Latin	(C0 Controls)	Cc	Clause 15	
0020		SPACE	Zs	Listed in 20.1	
007F		DELETE	Cc	Clause 15	
0080-009F	p425; C1 Controls and Latin-1 Supplement	(C1 Controls)	Cc	Clause 15	
00A0		NO-BREAK SPACE	Zs	Listed in 20.1	
00AD		SOFT HYPHEN	Cf	Listed in 20.3; Annex F1.1	
034F	p452; Combining Diacritical marks	COMBINING GRAPHEME JOINER	Mn	Listed in 20.3; Annex F1.1	
0600	p473; Arabic	ARABIC NUMBER SIGN	Cf	Listed in 20.3; Annex F.5	
0601		ARABIC SIGN SANAH	Cf	Listed in 20.3; Annex F.5	
0602		ARABIC FOOTNOTE MARKER	Cf	Listed in 20.3; Annex F.5	
0603		ARABIC SIGN SAFHA	Cf	Listed in 20.3	Not in Annex F
06DD		ARABIC END OF AYAH	Cf	Listed in 20.3; Annex F.5	
070F	p479; Syriac	SYRIAC ABBREVIATION MARK	Cf	Listed in 20.3; Annex F.5	
0F0C	p519; Tibetan	TIBETAN MARK DELIMITER TSHEG BSTAR	Po		???
115F	p528; Hangul Jamo	HANGUL CHOSEONG FILLER	Lo		???

Code Positions	p# from V4 book; Block name	Character name	Gen. Cat. ****	From Joint Edition Text	Candidate fro AFC list / annex F
1160		HANGUL JUNGSEONG FILLER	Lo		???
17B4	p560; Khmer	KHMER VOWEL INHERENT AQ	Cf	Notes in Annex P	???
17B5		KHMER VOWEL INHERENT AA	Cf	Notes in Annex P	???
180B	p564; Mongolian	MONGOLIAN FREE VARIATION SELECTOR ONE	Mn	Mentioned in clause 20.4; defined list of variants using VS-1 is tabulated in clause 20.4	Not called an AFC
180C		MONGOLIAN FREE VARIATION SELECTOR TWO	Mn	Mentioned in clause 20.4; defined list of variants using VS-1 is tabulated in clause 20.4	Not called an AFC
180D		MONGOLIAN FREE VARIATION SELECTOR THREE	Mn	Mentioned in clause 20.4; defined list of variants using VS-1 is tabulated in clause 20.4	Not called an AFC
180E		MONGOLIAN VOWEL SEPARATOR	Zs	Listed in 20.3; Annex F.2.5	
2000	p591; General Punctuation	EN QUAD	Zs	Listed in 20.1	
2001		EM QUAD	Zs	Listed in 20.1	
2002		EN SPACE	Zs	Listed in 20.1	
2003		EM SPACE	Zs	Listed in 20.1	
2004		THREE-PER-EM SPACE	Zs	Listed in 20.1	
2005		FOUR-PER-EM SPACE	Zs	Listed in 20.1	
2006		SIX-PER-EM SPACE	Zs	Listed in 20.1	
2007		FIGURE SPACE	Zs	Listed in 20.1	
2008		PUNCTUATION SPACE	Zs	Listed in 20.1	
2009		THIN SPACE	Zs	Listed in 20.1	
200A		HAIR SPACE	Zs	Listed in 20.1	
200B		ZERO WIDTH SPACE	Zs	Listed in 20.3; Annex F1.1; collection 200	
200C		ZERO WIDTH NON-JOINER	Cf	Listed in 20.3; Annex F1.1; collection 200	
200D		ZERO WIDTH JOINER	Cf	Listed in 20.3; Annex F1.1; collection 200	
200E		LEFT-TO-RIGHT MARK	Cf	Listed in 20.3; Annex F.1.3; collection 202	
200F		RIGHT-TO-LEFT MARK	Cf	Listed in 20.3; Annex F.1.3; collection 202	
2011		NON-BREAKING HYPHEN	Pd		???
2028		LINE SEPARATOR	Zl	Listed in 20.3; Annex F.1.2	
2029		PARAGRAPH SEPARATOR	Zp	Listed in 20.3; Annex F.1.2	
202A		LEFT-TO-RIGHT EMBEDDING	Cf	Listed in 20.3; Annex F.1.3; collection 203	
202B		RIGHT-TO-LEFT EMBEDDING	Cf	Listed in 20.3; Annex F.1.3; collection 203	
202C		POP DIRECTIONAL FORMATTING	Cf	Listed in 20.3; Annex F.1.3; collection 203	
202D		LEFT-TO-RIGHT OVERRIDE	Cf	Listed in 20.3; Annex F.1.3; collection 203	

Code Positions	p# from V4 book; Block name	Character name	Gen. Cat. ****	From Joint Edition Text	Candidate fro AFC list / annex F
202E		RIGHT-TO-LEFT OVERRIDE	Cf	Listed in 20.3; Annex F.1.3; collection 203	
202F		NARROW NO-BREAK SPACE	Zs	Listed in 20.3; Annex F.1.4	
205F		MEDIUM MATHEMATICAL SPACE	Zs		???
2060		WORD JOINER	Cf	Listed in 20.3; clause 8; Annex F1.1	
2061		FUNCTION APPLICATION	Cf	clause 8	???
2062		INVISIBLE TIMES	Cf	clause 8	???
2063		INVISIBLE SEPARATOR	Cf	clause 8	???
206A		INHIBIT SYMMETRIC SWAPPING	Cf	Listed in 20.3; clause 8; Annex F2.2; collection 205	
206B		ACTIVATE SYMMETRIC SWAPPING	Cf	Listed in 20.3; clause 8; Annex F2.2; collection 205	
206C		INHIBIT ARABIC FORM SHAPING	Cf	Listed in 20.3; clause 8; Annex F2.3; collection 205	
206D		ACTIVATE ARABIC FORM SHAPING	Cf	Listed in 20.3; clause 8; Annex F2.3; collection 205	
206E		NATIONAL DIGIT SHAPES	Cf	Listed in 20.3; clause 8; Annex F2.4; collection 206	
206F		NOMINAL DIGIT SHAPES	Cf	Listed in 20.3; clause 8; Annex F2.4; collection 206	
2FF0	p680; Ideographic Description Characters (visible displayable characters)	IDEOGRAPHIC DESCRIPTION CHARACTER LEFT TO RIGHT	So	Listed in 20.3; Annex F.3.2; collection 207	
2FF1		IDEOGRAPHIC DESCRIPTION CHARACTER ABOVE TO BELOW	So	Listed in 20.3; Annex F.3.2; collection 207	
2FF2		IDEOGRAPHIC DESCRIPTION CHARACTER LEFT TO MIDDLE AND RIGHT	So	Listed in 20.3; Annex F.3.2; collection 207	
2FF3		IDEOGRAPHIC DESCRIPTION CHARACTER ABOVE TO MIDDLE AND BELOW	So	Listed in 20.3; Annex F.3.2; collection 207	
2FF4		IDEOGRAPHIC DESCRIPTION CHARACTER FULL SURROUND	So	Listed in 20.3; Annex F.3.2; collection 207	
2FF5		IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM ABOVE	So	Listed in 20.3; Annex F.3.2; collection 207	
2FF6		IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM BELOW	So	Listed in 20.3; Annex F.3.2; collection 207	
2FF7		IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LEFT	So	Listed in 20.3; Annex F.3.2; collection 207	
2FF8		IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM UPPER LEFT	So	Listed in 20.3; Annex F.3.2; collection 207	

Code Positions	p# from V4 book; Block name	Character name	Gen. Cat. ****	From Joint Edition Text	Candidate for AFC list / annex F
2FF9		IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM UPPER RIGHT	So	Listed in 20.3; Annex F.3.2; collection 207	
2FFA		IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LOWER LEFT	So	Listed in 20.3; Annex F.3.2; collection 207	
2FFB		IDEOGRAPHIC DESCRIPTION CHARACTER OVERLAID	So	Listed in 20.3; Annex F.3.2; collection 207	
3000	p681; CJK Symbols and Punctuation	IDEOGRAPHIC SPACE	Zs	Listed in 20.1	
303E		IDEOGRAPHIC VARIATION INDICATOR	So		??? Not called an AFC
3164	p691; Hangul Compatibility Jamo	HANGUL FILLER	Lo	Listed in 20.3; Annex F.2.1; collection 204	
FE00- FE0F	p919; Variation Selectors	VARIATION SELECTOR-1 – VARIATION SELECTOR-16	Mn	Mentioned in clause 20.4; defined list of variants using VS-1 is tabulated in clause 20.4	
FEFF	p925; (BOM)	ZERO WIDTH NO-BREAK SPACE	Cf	Listed in 20.3; Annex F1.1; collection 200	
FFA0	p930; Halfwidth and Fullwidth Forms	HALFWIDTH HANGUL FILLER	Lo	Listed in 20.3; Annex F.2.1; collection 204	
FFF9	p936; Specials	INTERLINEAR ANNOTATION ANCHOR	Cf	Listed in 20.3; clause 8; Annex F.4	
FFFA		INTERLINEAR ANNOTATION SEPARATOR	Cf	Listed in 20.3; clause 8; Annex F.4	
FFFB		INTERLINEAR ANNOTATION TERMINATOR	Cf	Listed in 20.3; clause 8; Annex F.4	
FFFC		OBJECT REPLACEMENT CHARACTER	So	clause 8; mentioned in Appendix A.6	???
1D159	p959; Musical Symbols	MUSICAL SYMBOL NULL NOTEHEAD	So	Listed in clause 20.5; Annex U.2	???
1D173		MUSICAL SYMBOL BEGIN BEAM	Cf	Listed in clause 20.5; Annex U.2	???
1D174		MUSICAL SYMBOL END BEAM	Cf	Listed in clause 20.5; Annex U.2	???
1D175		MUSICAL SYMBOL BEGIN TIE	Cf	Listed in clause 20.5; Annex U.2	???
1D176		MUSICAL SYMBOL END TIE	Cf	Listed in clause 20.5; Annex U.2	???
1D177		MUSICAL SYMBOL BEGIN SLUR	Cf	Listed in clause 20.5; Annex U.2	???
1D178		MUSICAL SYMBOL END SLUR	Cf	Listed in clause 20.5; Annex U.2	???
1D179		MUSICAL SYMBOL BEGIN PHRASE	Cf	Listed in clause 20.5; Annex U.2	???
1D17A		MUSICAL SYMBOL END PHRASE	Cf	Listed in clause 20.5; Annex U.2	???
E0001	p1181; Tags	LANGUAGE TAG	Cf	clause 32 and clause 8; collection 3002 in Annex A; Annex T; collection 3002	???

Code Positions	p# from V4 book; Block name	Character name	Gen. Cat. ****	From Joint Edition Text	Candidate fro AFC list / annex F
E0020- E007F		TAG SPACE – CANCEL TAG	Cf	clause 32 and clause 8; Annex T; collection 3002	???
E0100- E01EF	p 1183; Variation Selectors Supplement	VARIATION SELECTOR-17 — VARIATION SELECTOR-256	Mn	clause 32 and clause 8; Mentioned in clause 20.4; collection 3002	???