**ISO/IEC JTC 1/SC 2 N 3354**
**Date: 1999-09-02**

ISO/IEC JTC 1/SC 2

CODED CHARACTER SETS

SECRETARIAT: JAPAN (JISC)

**DOC TYPE:** National Body Contribution

**TITLE:** National Body Comments on SC 2 N 3331, ISO/IEC 10646-1, Universal Multiple-Octet Coded Character Set (UCS) -- Part 1: Architecture and Basic Multilingual Plane, Second Edition text (Consolidation of ISO/IEC 10646-1: 1993, Amd 1 to 31, and Cor 1 to 3)

**SOURCE:** National Bodies of Japan and Republic of Korea

**PROJECT:** JTC 1.02.18.01

**STATUS:** This document is forwarded to WG 2 for consideration.

**ACTION ID:** ACT

**DUE DATE:**

**DISTRIBUTION:** P, O and L Members of ISO/IEC JTC 1/SC 2
WG Conveners and Secretariats
Secretariat, ISO/IEC JTC 1
ISO/IEC ITTF

**NO. OF PAGES:** 11

**ACCESS LEVEL:** Def

**WEB ISSUE #:** 061

Contact: Secretariat ISO/IEC JTC 1/SC 2 - Toshiko KIMURA
IPSJ/ITSCJ (Information Processing Society of Japan/Information Technology Standards Commission of Japan)*
Room 308-3, Kikai-Shinko-Kaikan Bldg., 3-5-8, Shiba-Koen, Minato-ku, Tokyo 105-0011 JAPAN
Tel: +81 3 3431 2808; Fax: +81 3 3431 6493; E-mail: kimura@itscj.ipsj.or.jp
*A Standard Organization accredited by JISC

# Japan's Comments on SC 2 N 3331

Attached please find comments on ISO/IEC JTC1 SC2 WG2 N2005 (WD 10646-1 edition-2).
Please note that Japan is in process of revising JIS version of ISO/IEC 10646-1 (JIS X 0221:95) to synchronize with the 2nd edition.
While the translating process, (even English version is not completed yet), JNB has had many comments on the N2005. Some of them might be seen as very significant technical comments which might need another amendment via balloting process. However, this paper includes all of those comments as well as minor editorial comments. Even though it is not right thing as a formality, to make better international standard available for the user, JNB is submitting all comments in this document. Editor and WG2 have a right to categolize the comments in to two, one should be adapted at time of the edition 2 publication and another for the later amendment(s) for the 2nd edition.

--------attachment---------

JPN-#1, Editorial
Clause 4.2
The second sentence of "4.2 block" "A block cannot overlap another block" should be rephrased to "A block does not overlap another block".

---

JPN-#2, Minor Technical
Clause 4.33
On "4.33 RC-element", the wording looks unclear in the following two points:

- RC-elements are used only in UCS-2 and UTF-16 representation and never be used in UCS-4 and UTF-8 representation, but the fact is not obvious in current text.

- Current definition is misleading that any cells (including those outside of BMP) have a corresponding RC-element, although both UCS-2 and UTF-16 only require RC-elements corresponding to cells in BMP.

Japan suggests the following text for clarification:

"A two-octet sequence comprising the R-octet and the C-octet from the four-octet canonical form (see 6.2) of a cell in BMP. RC-elements are used in UCS-2 representation and UTF-16 representation of this coded character set."

---

JPN-#3, Editorial
clause 4.33
The "4.33 RC-element" should be renumbered to "4.32 RC-element".

---

JPN-#4, Minor Technical

clause 4.23, 4.26
On "4.23 high-half zone", wording is unclear, since it is not obvious that wording following the semicolon only applies to UTF-16. Also, the word "reserved" is misleading, since the word may be interpreted as "never use it." (Users are allowed to use RC-elements from high-half zone to form a valid paired RC-element, of course.) Japan suggests to rephrase it as follows: "A set of cells to be used as the first of a pair of RC-elements which represents a character from a plane other than th BMP in UTF-16 (see Annex C)."

Same thing for "4.26 low-half zone".

---

JPN-#5, Editorial
clause 5
The last two paragraphs in "5 General structure of the UCS" looks strange. They should be rephrased as follows:

> "Two UCS Transformation Formats, UTF-16 and UTF-8, are specified in Annex C and D, respectively.
>
> UTF-16 can be used to represent ...
>
> UTF-8 can be used to transmit ..."

---

JPN-#6, Minor Technical
clause 6.5
On "6.5 Identifiers for characters", several terms are used for a same thing; they are: "identifier for character", "short identifier", character identifier", and just "identifier". A single term should be used consistently. Japan suggests "short identifier". Moreover, the "short identifier" is a unique notion, so it should be listed (and defined) in "4 Definitions".

---

JPN-#7, Minor Technical
clause 6.5
In the first paragraph in "6.5 Identifiers for characters", the last sentence of the paragraph describes short identifiers in translation of the standard text. Japan believes the sentence should be written as a note or be deleted, for the following two reasons:

- Issues in translation of an IS are a matter of standard-development process and have nothing with users of the standard. It is not appropriate for the normative text.

- When compared to "6.4 Naming of characters", it is unnatural and misleading to describe relationship with translated version only in 6.5.

---

JPN-#8, Editorial
clause 6.5

On the 8th paragraph in "6.5 Identifiers for characters", the two capitalized words "CAPITAL" and "SMALL" should be written normally (i.e., using lower case), since they are normal English words and not character names.

---

JPN-#9, Minor Technical
clause 13
In the first paragraph in "13 Coded representation forms of the UCS", rephrase "ISO/IEC 10646 provides two alternative forms" to "ISO/IEC 10646 provides four alternative forms", since we now have four normative forms. Also add the following at the end of the paragraph: "Two of those forms, UCS-2 and UCS-4, are defined in this clause; other two, UTF-16 and UTF-8, are defined in Annex C and D, respectively."

---

JPN-#10, Editorial
clause 15
On "15 Use of control functions with the UCS", it is unclear how to pad (and when not to pad) controls in UTF-8 and UTF-16, when reading clause 15. Japan suggests to add ", Annex C and Annex D" after "see clause 13" at the end of the first sentence of the second paragraph in clause 15.

---

JPN-#11, Minor Technical
clause 16.2
On "16.2 Identification of UCS coded representation form with implementation level", it is unclear that there are more designation sequences for UTF-16 and UTF-8. Japan suggests to rephrase the last part of the first paragraph of 16.2 "... shall be by a designation sequence chosen from the following list:" to "... shall be by a designation sequence chosen either from the list specified in C.5, from the list specified in D.6, or from the following list:"

---

JPN-#12, Major Technical
clause 16.5
The last paragraph in "16.5 Identification of the coding system of ISO/IEC 2022" should be removed, since the condition for which the paragraph defines the bit combinations is impossible (i.e., the escape sequence "return from UCS to 2022" is used in CC-data-element conforming to 2022).

---

JPN-#13, Editorial
clause 25.2
In the first example (Row 0B Tamil) in "25.2 Features of Indic alphabetic scripts", "... appears is if ..." should be corrected to

"... appears as if ...".

---

JPN-#14, Minor Technical
clause 25.2
Current description in "25.2 Features of Indic alphabetic scripts" is misleading. Japan has the following concern:

- Current text introduces the issue as "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." This statement gives a wrong impression to users that the issue is strictly on graphic symbols (or, rendering.) The standard text should be clear on the point that the issue is about the _logical_ characters and is primarily independent from their graphic representation.

- Statement like "appear to be formed as" is very ambiguous. Readers unfamiliar with Indic scripts could think of a character "appears to be formed" from unrelated characters. On the other hand, it is almost obvious for an Indic-script-familiar person which vowel is a combination of which vowels. We should be honest about the fact that readers are expected to learn features of Indic scripts by themselves, and the standard only describes only how the feature is handled in UCS.

- The requirement described in the last paragraph is misleading, since it only says about Level 1 and Level 2, and the _rule_ described using "shall" looks like a normal handling. It will be much clearer if we write the same thing upside-down; i.e., a special equivalence handling is allowed only on Level 3.

- A term "unique-spelling rule" is used in this clause as well as clause 14. However, there is no explicit definition what is the "unique-spelling rule." Current text just says "when this rule applies, ..." We need an explicit statement what is the rule.

Japan suggests the following wording for the first paragraph:

> In Indic scripts, a combination of two characters (typically but not necessarily two vowel signs) may be coded as a separate character occupying another code position.

and the following for the last two paragraphs:

> In rendering, such a character is expected to be displayed equivalently to the case of a series of two characters. This behaviour is similar to the case of composite sequences.
>
> Since the character and the corresponding sequence of two other characters are coded in separate positions, except for the case of Level 3, they shall make those characters as separate entities to a user when those particular characters are supported. This character handling is called "unique-spelling" rule.
>
> In level 3, however, such a character and the corresponding

series of characters may be regarded as equivalent entity.
(I.e., "unique-spelling" rule need not apply in Level 3.)

---

JPN-#15, Editorial
Annex A A.1
"Note 2" in A.1 has the following problems:

- A normative text in the second paragraph in A.1 has an explicit reference to the NOTE 2.

- Notes 1 and 2 are isolated by normative texts (or, in other words, notes for different normative texts are numbered sequentially), which violates the Drafting Rules of International Standard.

Japan suggests:

- to remove the second paragraph in A.1, and

- to move NOTE 1 immediately before NOTE 2.

---

JPN-#16, Editorial
Annex A A.1
On A.1, there are two entries, 57 and 58, which is described as "[deleted at Amd.5]", as a part of normative text. This type of informative explanation is inappropriate for a normative text.

Also, A.1 includes another entry "[299 BMP FIRST EDITION] see A.3*".
The meaning of the surrounding angular blacket is unclear.

Japan suggests to isolate normative texts and informative texts, rewriting them as follows:

- Replace entries for 57 and 58 as "57 (This collection number shall not be used+.)" and "58 (This collection number shall not be used+.)", where "+" is a footnote daggar.

- Add the following footnote:

  +) Collection numbers 57 and 58 were specified in the first edition of ISO/IEC 10646-1, and deleted by its Amendment 5. Use of collection numbers 57 and/or 58 is not now in conformity with this International Standard.

- Replace entry for 299 as "299 (This collection number shall not be used++.)", where "++" is a footnote double dagger.

- Add the following footnote:

  ++) See A.3.2.

---

JPN-#17, Editorial
Annex C
On "Annex C (normative) Transformation format for 16 planes of Group 00 (UTF-16)", notes have consequtive numbers throughout the Annex, violating the Drafting Rule of International Standard. A note should not have any number, when only one note appears after a paragraph. Hence, "NOTE 1" in C.3 and "NOTE 2" in C.6 should be just "NOTE".

---

JPN-#18, Editorial
Annex C C.6
The example in C.6 is inappropriate and could mislead the reader rather than help correct understanding. It has the following problems:

- There are no explanation what a string surrounded by "<" and ">" which means; it is hard to find a <xxx> represents an RC-element which does not correspond to ASCII graphic character. We should use four digit hexadecimal notation to explicitly write RC-element (with some explanatory _rendaring_ attached.)

- Character string "<box>" should be printed as a real box character, insread of sequence of ASCII "<", "b", "o", "x", ">". It could mislead the reader from the point this example tries to explain.

- Use of a hieroglyph as an example of a character in several places between plane 1 to 16 is not a good practice, since exact allocation of hieroglyph will not yet have been fixed at the time of publication of the Edition 2. We should instead use more stable characters which will be available to the public.

- The example uses several uncommon (to typical readers) terms like Latin-1 or glyphs without strong reason to do so. Those sentences should be rephrased with plain wording.

Hence, the example should be re-wrote something like:

> Example: Assume a device receives the following sequence of RC-elements:
>
> 0047 0072 0065 0065 006B 0020 03B1 0020 0061 006E 0064
> G r e e k + a n d
>
> 0020 0045 0074 0072 0075 0073 0063 0061 006E D800 002E
> E t r u s c a n .
>
> If the device only handles the BASIC LATIN subset and uses a box (*) to indicate unsupported characters, it would display:
>
> Greek * and Entruscan *.

NOTE - A character ETRUSCAN LETTER A has a code position 00010300, which is specified in Part 2 of this International Standard. A correct UTF-16 representation of this character is D800 DE00.

where + and * should be printed using real _glyph_ for a greek alpha and a box.

Similar comments applies to examples in C.7; "<" and ">" enclosed string should be avoided, and use a stable character instead of Phenicia as an example.

---

JPN-#19, Editorial
Annex D
On "Annex D (normative) UCS Transformation Format 8 (UTF-8)", notes have consequtive numbers throughout the Annex, violating the Drafting Rule of International Standard. They should have a paragraph-to-paragraph numbering. Hence, "NOTE 1" and "NOTE 2" in D.2 and "NOTE 5" in D.5 should be just "NOTE", and "NOTE 3" and "NOTE 4" in D.4 should be "NOTE 1" and "NOTE 2" respectively.

---

JPN-#20, Major Technical
Annex D
On "Annex D (normative) UCS Transformation Format 8 (UTF-8)", it is unclear how we can use C1 controls in UTF-8. Several interpretations are possible.

- Since the standard doesn't specify any direct way to represent C1 controls in UTF-8, one can interpret it as an implicit prohibition of use of C1 controls in UTF-8 other than ESC Fe representation. (C1 controls 80 to 9F are mapped to 1B 80 to 1B 9F, in this case.)

- We can simply apply the mapping rule specified in D.4. That is, pad a C1 control according to the caluse 15 to get four octet sequence to be used in CC-data-elements conforming to UCS-4, and applies rules specified in D.4 to the four octet sequence as if it is a valid UCS-4 representation of a UCS character. (C1 controls 80 to 9F are mapped to C2 80 to C2 9F, in this case.)

- We could interpret that we had to apply padding rule specified in caluse 15 directly. The clause requires us to pad controls with zero octets up to "the number of octets in the adpoted form." The number of octets in UTF-8 is unclear, but it apparent to be a number between 1 and 6, inclusive. C1 controls 80 to 9F could be mapped to 80 to 9F, 00 80 to 00 9F, 00 00 80 to 00 00 9F, ..., or 00 00 00 00 00 80 to 00 00 00 00 00 9F, in this interpretation.

At this time, Japan has no strong opinion about which interpretation should be adopted, but one particular interpretation should be agreed in WG2 and some appropriate changes should be applied to Annex D.

---

JPN-#21, Editorial
Annex L

The second paragraph of "Rule 3" in "Annex L (informative) Character naming guidelines" is misleading. The paragraph gives a wrong impression that the * (asterisk) is a part of the official name and that the asterisk is omitted only in Annex G. (It sounds like anybody must retain the asterisk whenever uses a name outside of Annex G.)

Japan suggests to replace the second paragraph of "Rule 3" with the following note:

> NOTE. The name of a character may be followed by a single * symbol, when used in this International Standard. This indicates that additional information on the character appears in Annex P. Any symbol * is not a part of the name of a character.

---

JPN-#22, Minor Technical
Annex N N.3
On the fourth paragraph in "N.3 Identification of ASN.1 character transfer syntaxes", two identifiers "UTF16-form" and "UTF8-form" listed to show object identifiers violate ASN.1 syntax, since their identifier parts should begin with lowercase letter.

Japan suggest to correct their identifiers to "uTF16-form" and "uTF8-form" respectively.

---

JPN-#23, Editorial
Annex P
After the third paragraph in Annex P, something which looks like an unnumbered caluse title ("Group 00, Plane 00 (BMP)") exists. It has two problems.

- All concrete characters described in 10646-1 is in BMP, (since Part 1 only covers architecture and BMP,) so it is obvious that all characters explained in Annex P are in BMP. The text "Group 00, Plane 00" is redundant.

- The style the line is written violates the Drafting Rules of International Standard.

Japan suggests just to remove the line.

---

JPN-#24, Editorial
Annex P
On caluses in Annex P, each caluse starts with a code position followed by a character name. This style has following two problems:

- The second paragraph in Annex P says "Each entry in this annex consists of the name of a character and its code position". It implies the name comes before code position.

- Clauses in Annex F, which have similar style as Annex P, list

characters as their names followed by their code positions in parentheses. We should use consistent style.

Japan suggests to use a same style as Annex F, i.e., to start each clause with a character name followed by a parenthesized code position, a collon, then its explanation (in a same paragraph.)

---

JPN-#25, Editorial
Annex P and other places
On several parts, e.g. 26.2, Annex P, Annex Q, words "hex" or "hexadecimal" is prefixed before a hexadecimal notation for a UCS code position. 10646-1 explicitly specifies (in 6.2) "The value of any octet shall be represented by two hexadecimal digits", so prefixing "hex" or "hexadecimal" is redundant. Moreover, such redundant notation may be misleading when some hexadecimal value consists only of decimal digits be interpreted as a decimal number. Those redundant "hex" and "hexadecimal" prefix should be removed.

---

JPN-#26 Minor technical
Annex P
FA1F, FA23; Add character symbols of each national body.

---

JPN-#27 Editorial
Annex P
FA1F 7th line
Change from IDEO-GRAPHS to IDEOGRAPHS
Rationale: This might be typo

---

JPN-#28 Minor technical
Annex P
FFE3 FULLWIDTH MACRON 2nd line
Change From: It may also be used as
To: It is also used as
Rationale: This is actually naming mistake. It should be OVERLINE as a reality.

---end of comments------

<Korean Comments>

Subject: Submitting Korean comments on SC2N3331(ISO/IEC 10646-1, Part 1, Second Edition)

After we checked ISO/IEC 10646-1, Part 1, Second Edition (SC2 N3331) for review and comments, we would like to mention two points.

1. On page 16, in Table 1: Elements of Hangul syllable names and annotations, index number 0 is used for G in I string column (i.e., syllable initial G), whereas index number 1 is used for the same G in F string column. However, the same number (usually 1) is used for both syllable-initial and syllable-final G in most documents. In addition, index number 1 is usually used for A in P string column.

Although there is no technical problem, using different index numbers in different documents could cause unnecessary consufion. Considering that ISO/IEC 10646 will have a large impact, we suggest that we prevent this kind of unnecessary confusion.

The same comment can be said of the Annotation elements: e.g., syllable-initial k and syllable-final k.

We suggest that

1) we change the starting index number for I string and G string from 0 to 1 for both Syllable name elements and Annotation elements; and
2) we change other necessary statemetns/formula/etc due to change 1)

2. Note 2 at the end of Section 25.1

Editorial change regarding the usage of Bangjeom letters is included at the end of Section 25.1. Since we requested that an "ordinary, not combining" Bangjeom letters be included in 10646-1, we suggest that this editorial comment be deleted. We further request that WG2 discuss regarding ordinary Bangjeom letters.

3. Compact name table for Hangul syllables

As Mr. Bruce Paterson suggested, we suggest that we add an Annex (informative) for 10646-1 2nd edition to list the Hangul syllable names in a very compact format (11 pages).