ISO/IEC JTC 1/SC 2 N 3332/WG2 N 2057

Date: 1999-06-22

ISO/IEC JTC 1/SC 2

CODED CHARACTER SETS

SECRETARIAT: JAPAN (JISC)

DOC TYPE: Other document

TITLE: National Body Comments on SC 2 N 3297, WD 10646-2, Universal Multiple-Octet Coded Character Set (UCS) -- Part 2: Secondary Multilingual Plane for scripts and symbols Supplementary Plane for CJK Unified Ideographs General Purpose Plane

SOURCE: National Bodies of Japan, Sweden, UK and USA

PROJECT: JTC 1.02.18.02

STATUS: This document is forwarded to the project editor and 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: 7

ACCESS LEVEL: Defined

WEB ISSUE #: 055

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

ISO/IEC JTC 1/SC 2 N

Date: **1999-06-15**

ISO/IEC JTC/1 SC/2
Coded Character Set
Secretariat: Japan (JISC)

Doc. Type: National body comments

Title: **Comments on SC2 WG2 N2012R (ISO/IEC WD 10646-2)**

Source: **Japan**
Project: JTC1 02.18.02
Status: For review by WG2
Date: 1999-06-15
Distribution: SC2
Reference: SC2 WG2 N2012R
Medium:

The document SC2 WG2 N2012R (ISO/IEC WD 10646-2) is not mature enough to forward as a CD of the JTC1 standard. It is still, and almost good as, an explanation document of the basic idea of the ISO/IEC 10646-2. But not yet as WD of the international standard of ISO and/or IEC. Therefore, for the review for CD, Japan request one more WD cycle concerning following comments.

1. Tagging character and conformance clause.
   The WD says that the part-2 will share the comformance clause with the part-1. The conformance clause of the part-1 is defined to aim a graphic characters with some functionality (if any). Therefore the characters defined in the part-2 should have only graphic shape, graphic shape with clearly defined functionality or clearly defined functionality with invisible (non printable) graphic shape.
   The tagging characters defined in plane-14 are none of above. There is no clear definition of the functionality with the characters. There is only one sample in case of the “use for language tagging” in annex C. This is not a way that JTC1 standard should be written. Existing conformance clause does not work for the example.

   There are four possibilities to resolve this fundamental problem.
   1. Rewrite conformance clause to fit the existing text. (need addition of some explanation text for tagging character in general)
   2. Add new clause to define “what”s tagging character?”.
   3. Change “generic tagging character” to “language tagging character” in this case, sample use in annex C should be moved to main text with proper modification.
   4. Withdraw the “tagging characters”.

   Japan recommend 3 or 2 above. (3 with higher priority)
   This does not automatically mean that Japan agree with the idea of the (language) tagging character. Japan still hold the reservation a right to comment to the idea after the new text is available.

   Note 1: Some of the characters in the part-1 are also needed to add an explanation for conformant reason such as OBJECT REPLACEMENT CHARACTER.
   Note 2: Is Clause 2 Note 1 applicable for the TAG CHARACTER?

2. Annex C

2-1. Remove 1st line of C.2. it is unnecessary. Start with “ In order to….”.

2-2. Add a description of “tag identification characters” in main text and sample of “language tag identification Character” in annex C. There is no clear explanation “how to identify the tagging characters for specific tagging purpose (even in case of language tagging)”.

3. Several important clauses and annexes are missing as a JTC1 standard.

3-1. Part-1 clause 27 (in 2nd ed.) equivalent. Even though the printed format is not finalized yet, there is some new text are needed to express the relation with the clause 26 of part-1 (the text from the part-1 is not direct applicable as it is). So partical text to be provided for CD review.

3-2 Sample of character name table (may be, source table) for plane-2 should be added. Remember that the character source information has been removed form the code table of the plane-2 per request of WG2 and editor. Editor might have an idea to handle it. (IRG will provide source information only)

3-3. Annex for combining characters At least, in annex B (a list of combining character), there should be a text which explains the relation with the part-1, and may be, there might be some reflections on the part-1 (implementation levels).

3-4 Annexes for character sources and additional information on characters. Since the part-2 is for rerely used characters, those information is fur more important than that of the part-1.

3-5. May be, same kind of review are needed for other annexes and clauses in part-1. For exsample, Use of same unification rule is mentioned already, but direct reference to the annex T (of 2nd ed.) of the part-1 is questionable.

3-6. Sub-setting: Need to have some concern (and proposal) on sub-setting and default set. When ISO/IEC 10646 conformant is claimed, is that mean all planes are available? Can we say ISO/IEC 10646-1 conformant? Or always needed to say BMP subsetting? How about in case additon of characters are done by amendment(s)?

3-7 Add ISO/IEC 10646-1 to the clause 3 normative reference.

3-8 If graphic characters in GPP “MAY NOT” have a visual representation. Clearly state ” which shall have” and “which shall not have” a visual representation. At least clause for reading a code tables is needed.

3-9 Annex A: Add BOCK and COLLECTION for CJK EXT-B.

3-10 Do MUSICAL NOTEs have only graphic shape? If those (or some of those) do have a semantics along with graphic shapes. Describe them in bew aneex.

3-11 State that there is no such a characters like clause 20 of the part-1 in this part-2.

4. Other comments

4-1. Name of plane: General Purpose Plane (GPP) is misleading. Consider to change the name. (such as Special Characters Plane (SCP))

4-2. Clause 4. 1st line. Change “Part 1” to “ ISO/IEC 10646 Part-1” or “Part 1 of this International Standard”

4-3. Clause 6. 2nd paragraph. 1st line. : Change “special plane” to “separated plane” (or some thing else. The plane reserved for CJK is not “special”. (another possibility might be “specific plane fot CJK unified ideograph supplementation are reserved)

4-4. Review whether if additional Esc sequence is needed or not on clause 17.2 of the part-1. (in relation with sub-stting and super-setting)

5. Other ideas

5-1. Consider to use same annex number as the part-1 if possible.

5-2. Add an informative annex about UTF-16 access areas (planes), such that the user of the standard might have better view of the part-2. Or add special note in the clause 1 SCOPE to talk about UTF-16 planes.

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

The Swedish NB is of the opinion that the "tag characters" shall not be standardised, at least not at present. For further comments see below.

Comment from Sweden.

The "general purpose plane" (plane 14) is by this WD used for "tag characters". These are essentially yet another copy of the ASCII characters in the UCS plus two "tag syntax" related characters (LANGUAGE TAG and CANCEL TAG). We object to the plane 14 characters for several reasons:

a. Language tags are already included in several 'higher level protocols'. If language tagging is needed, one can use either the mechanisms already provided in XML/SGML/HTML, or use some similar scheme for language tagging, possibly simpler, possibly adapted to a different kind of markup. (Note that the plane 14 characters are ill-suited for use with HTML/XML, and are likely to become disallowed in HTML/XML files.)

b. The plane 14 tag characters are made for a particular syntax for the language tags. What the syntax for language tags is, is clearly out of scope for 10646.

c. The plane 14 tag characters are limited to expressing the tag values in (shadow) ASCII. In a standard such as 10646 a limitation to use only ASCII (remapped) is very strange. Indeed, languages may have (informal or formal) identifications that include non-ASCII characters.

If anything like tag characters are to be acceptable from our point of view, then the character allocation for a special syntax only, must be generalised to allow any syntax (which one to use is out of scope for 10646), and the restriction to (shadow) only ASCII must also be removed. Any characters must be allowable to use as "tag characters", for language tagging or otherwise.

Our primary preference would be to simply remove the plane 14 tag characters from any further consideration, leaving language tagging only to markup (e.g. XML; or something much simpler). Note that XML/HTML files should never use the plane 14 language tags anyway.

Our secondary preference is to use a completely different scheme for this, and similar, kind of tagging that is general enough both in being able to completely move tag syntax considerations out of 10646, and to allow any characters (or rather 'shadow characters' for all other UCS characters) in tags. Possibilities for this include, but are not restricted to,

a) using all of plane 14 for a UTF-16 remap,
b) using the 256 first characters in plane 14 for a UTF-8 remap,
c) using a single "META" character code point, the use of which marks some nearby character as being a tag, or meta, character.

We understand that the tag characters have been proposed in order to get easily identifiable (language) tags, the identification to be done by just looking at character codes rather than parsing any markup that uses ordinary characters. However, it is very doubtful that this is really needed. Parsing simple language tagging expressed with ordinary characters (e.g. <svenska>Hej</svenska>, if something HTML-inspired is used) can be made simple enough.

For the application originally in mind for the plane 14 tag characters ('name'-'language tagged string value' pairs in Internet protocols) an even simpler approach can be used instead:

Using plane 14 tags:
attribute_name: <some plane 14 tagged string with tagseparated message in multiple languages>

Without plane 14 tags:
attribute_name,sv: <message in Swedish>
attribute_name,en_UK: <(same) message in UK English>
attribute_name,jp: <(same) message in Japanese>
attribute_name,zh_HK: <(same) message in traditional Chinese>

Plane 14 tag characters appears to serve no purpose being allocated in 10646. The functionality they offer can with preference be replaced by other language tagging methods that do not require any special character allocations in 10646.

Consequently it is proposed that sections 5.3, 8 and 9.3 and Annex C are deleted, as well as the parts of Annexes A and B that refer to the plane 14 characters."

UK Comments on WD 10646-2, Universal Multiple-Octet Coded Character Set (UCS) -- Part 2:
Secondary Multilingual Plane for scripts and symbols Supplementary Plane for CJK Unified Ideographs General Purpose Plane
(SC 2 N 3297)

Technical comments

Tables 4 and 5, Byzantine Musical Symbols.

1. The source document WG2 N 1582 identified, for each named character, the position upper/middle/lower to show the stripe within which each character should be rendered. This information should be retained in the WD, either as annotations for the character names, or by a marking in each glyph cell.

2. The Resolution M33.12 which approved this character set required that combining characters should be indicated. This information is missing from the WD, and must be obtained from the originator of the source document, since it does not appear in WG2 N1582.

Tables 6 and 7, Western Musical Symbols

3. Since many of the musical symbols in this set will normally be rendered superimposed on a section of STAFF, either the STAFF symbols D1CD - D1D2 should be specified as combining characters, or a separate set of COMBINING STAFF symbols should be provided.

4. Since the vertical positioning of many symbols on a staff must be under user control, and is of primary significance, a suitable formatting function should be provided to indicate the number of vertical steps, up or down relative to the "base" position of the staff, that a given symbol is intended to be rendered.

5. The STEM and TREMOLO symbols D196 - D19B will normally be combined with other symbols, so they should be specified as combining characters, to clarify whether they should appear in the data stream before or after the characters that they combine with.

Editorial comments

Clause 4, paragraph 2: replace "planes 01 to 15" by "planes 01 to 0F".
Clause 4, Note: replace "plane 16" by "plane 10", twice.
Clause 5.3, clause 8, Table 8 (both pages):
Replace "Plane 14" by "Plane 0E".

BSI
15 June 1999

Re: Comments on the working draft of ISO/IEC 10646 Part 2 (WG2 N 2012R2)
From: US
Date: 1999-06-11

The US requests that the following language be added to Part 2, end of clause 8 (General Purpose Plane), for compatibility with the Unicode Standard:

"To allow a greater degree of compatibility across versions of the standard, the ranges U-000E0000..U-000E1000 are reserved for future alternative format characters."

The US suggests that the following note be added after this clause.

Unassigned code points in these ranges should be ignored in normal processing and display.

The US requests that the following language be added to Part 2, as an amendment to Part 1, clause 8 (Basic Multilingual Plane):

To allow a greater degree of compatibility across versions of the standard, the ranges U-00002060..U-00002069 are reserved for future format characters.

The US suggests that the following note be added after this clause.

Unassigned code points in these ranges should be ignored in normal processing and display.

For information, the following text is being published in the Unicode Standard, Version 3.0.

Unassigned Characters

In practice, applications must deal with unassigned code points. This may occur, for example, when the application is handling text that originated on a system implementing a later release of Unicode with additional assigned characters. To work properly in implementations, unassigned code points must be given default properties as if they were characters, since various algorithms require properties to be assigned to every character in order to function at all. These properties are not uniform across all unassigned code points, since certain ranges of code points need different properties to maximize compatibility.

The Unicode Bidirectional Algorithm assigns directional properties based on the expected direction of characters to be added in the future. All unassigned code points in Hebrew, Arabic, Thaana, and Syriac blocks are given the bidirectional property R (right-to-left). These are the ranges 0590-05FF, FB1D-FB4F, 0600-07BF, FB50-FDFF, and FE70-FEFF. All other unassigned code points are given the bidirectional property L (left-to-right).

Normally, code points outside the repertoire of supported characters would be displayed with a fall-back glyph, such as a black box. However, format and control characters must not have visible glyphs (although they may have an effect on other characters in display). These characters are also ignored except with respect to specific, defined processes: for example, ZERO WIDTH NON-JOINER is ignored in collation. To allow a greater degree of compatibility across versions of the standard, the ranges 2060-2069 and 000E0000-000E1000 are reserved for future format and control characters. Unassigned code points in these ranges should be ignored in processing and display.