ISO/IEC JTC1/SC2/WG2 N 4521
Date: 2014-02-10

ISO/IEC JTC1/SC2/WG2
Coded Character Set
Secretariat: Japan (JISC)

Doc. Type: **Draft disposition of comments**

Title: **Draft disposition of comments on ISO/IEC DIS 10646 4th edition**

Source: **Michel Suignard (project editor)**
Project: JTC1.02.10646.00.00.00.04
Status: For review by WG2
Date: 2014-02-10
Distribution: WG2
Reference: SC2 N4307 WG2 N4493
Medium: Paper, PDF file

Comments were received from Ireland, Japan, UK, and USA. The following document is the proposed disposition of those comments. The disposition is organized per country.

Note – With some minor exceptions, the full content of the ballot comments have been included in this document to facilitate the reading. The dispositions are inserted in between these comments and are marked in **Underlined Bold Serif text,** *with explanatory text in italicized serif***.**

Ireland: Negative

Ireland disapproves the draft with the technical and editorial comments given below.
Acceptance of these comments and appropriate changes to the text will change our vote to approval.

Editorial comments

E1. Page 1254: Row 10C8: Old Hungarian.

After consultation with colleagues in Hungary, Ireland recognizes that modern orthographic use of a few of the Old Hungarian letters differs somewhat from that of the primary-source manuscripts. Ireland requests the addition of the following informative notes to clarify some modern usage:

𐳧 10CA7 OLD HUNGARIAN CAPITAL LETTER ENT
- *also used for Ant and Int*

𐳬 10CAC OLD HUNGARIAN CAPITAL LETTER NIKOLSBURG UE
- *= Ü*
- *also used for Ö*
- *used for U with double acute in Sándor Forrai's orthography*

𐳭 10CAD OLD HUNGARIAN CAPITAL LETTER RUDIMENTA UE
- *= Ü, U with double acute*
- *used for Ü in Sándor Forrai's orthography*

𐳷 10CE7 OLD HUNGARIAN SMALL LETTER ENT
- *also used for ant and int*

𐳼 10CEC OLD HUNGARIAN SMALL LETTER NIKOLSBURG UE
- *= ü*
- *also used for ö*
- *used for u with double acute in Sándor Forrai's orthography*

𐳽 10CED OLD HUNGARIAN SMALL LETTER RUDIMENTA UE
- *= ü, u with double acute*
- *used for ü in Sándor Forrai's orthography*:

Propose acceptance

E2. Page 1254: Row 10C8: Old Hungarian.

Ireland believes that the glyphs of two letters (four characters) should be modified slightly. 10C8B OLD HUNGARIAN CAPITAL LETTER EE and 10CCB OLD HUNGARIAN SMALL LETTER EE are the casing forms of a modern letter, devised in the 20th century, and their glyphs should be shorter, to harmonize with the preferences of modern users. The glyph for 10C93 OLD HUNGARIAN CAPITAL LETTER EK and 10CD3 OLD HUNGARIAN SMALL LETTER EK is drawn rather shorter than other letters in the Nikolsburg and Rudimenta manuscripts (see N4268R figures 4 and 5) but in the Bologna manuscript and elsewhere in the Rudimenta it is drawn more or less the same height as other letters (see N4268R figures 2 and 3). Modern users tend to prefer the larger letter, and as this shape is attested in the primary materials as well, Ireland requests that the glyphs of the four characters be changed to the glyphs as shown below, namely:

From this set of glyphs:  to this set of glyphs: 

Propose acceptance

A new font with these characters has been provided to the editor.

Japan: Positive with comments

Technical, and Editorial comments (noted as T or E)

E1. 2.2 and many others

First character of “Clause” is capitalized

Proposed change by Japan

Change to lower case character.

Propose non acceptance

Checking the ISO Directives Part2 ‘Rules for the structure and drafting of International Standards’ themselves, there are no clear indication of the preferred style and the text of the directives never refers to its own clauses, only sub-clauses. All references using non numeric references have title case on their names, such as ‘see Annex X’ or ‘see Example’. However the ISO Directives Part 1 text uses twice ‘see [] Clause x’.

T2. Sub-clause 23.2 – CJKsrc.txt file

In CJKsrc.txt file, UCS code point is described as the format of “U+hhhh” or “U+hhhhh”. This is not consistent with other UCS code point format “hhhh” or “hhhhh”, for example, in 22.4.)

Proposed change by Japan

Change the format to “hhhh” or “hhhhh” in CJKsrc.txt.

Propose Non acceptance

While it is true that this file uses a different format for UCS code point than other data file as noted by Japan, nothing prevents the usage of the U+hhhh or U+hhhhh which is an allowed syntax per sub-clause 6.5 ‘Short Identifiers for code points (UIDs)’. In this case, the syntax choice is determined by the desire to share the same format as the Unicode Han database (Unihan). A note could be added to that effect in the Table 5 providing the format details.

T3. Sub-clause 31.3 – Code Chart for CJK Extension E

(This is to confirm the resolution IRG M41.2 from IRG meeting #41 in Tokyo.)
As described in IRG N1957, there are glyph errors in CJK Extension E code table as for U+2C0D1 and U+2C7D3.

Proposed change by Japan

Make sure to correct the glyphs shown as IRGN1977 with reflecting the feedback to this document if any.

Propose acceptance

The change has already been incorporated in the working set for the next phase for publication.

E4. Annex A

The title of A.6.12 is “312 Unicode 6.3”.

Proposed change by Japan

Change to “313 Unicode 6.3”.

Accepted

This was a typo.

E5. Annex F

“sub-clause” has a dash, although other “subclause” does not.

Proposed change by Japan
Keep them consistent.

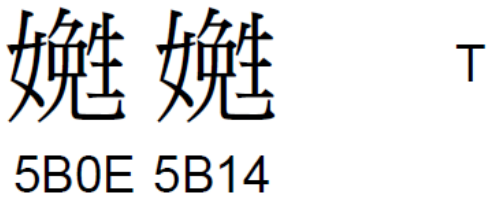
Propose acceptance

In the standard there is one occurrence of subclause in 2.3 and three occurrences of sub-clauses in Annex F. The term 'subclause' (without space or dash between sub and clause) is much rarer than sub-clause (ration of 1 to 30 in search engines) but in use in ISO Directives. The editor is open to preferences from experts.

T6. Annex S

(This is same comment as what JP commented at CD ballot of ISO/IEC 10646 4th edition.)

The following pair taken from S.3 is for a source separation caused by T source.



This figure should be taken from 2003 edition in order to preserve the original intention from the early days as follows.

Here, a shape with a dot is on 5B14, and that without is on 5B0E.

Proposed change by Japan
Use the figure from 2003 edition.

Propose non acceptance

First it is important to show the disposition of comment from the CD. The comment was not just about these 2 characters but about all of Annex S.3 and S.4.

Partially accepted

There are several considerations:

- *Many editions have been published since 2003 and no one has objected to the update of these two sub-clauses until now,*
- *The other sub-clauses of Annex S have already been reversed to pictures where it mattered,*
- *Unlike the other parts of Annex S, sub-clauses S.3 and S.4 contains code points and readers may be surprised that the shapes shown there do not correspond to the IRG source glyphs,*
- *Glyph outlines look much better than pictures,*
- *The modifications to S.3 and S.4 were done before the multi-column format for CJK was done and before the various IRG sources fonts were available to the editor. Commercial fonts were used which are sometimes quite different from the official sources. Now it is possible to use the IRG source glyphs as shown in the chart pages.*

Based on these considerations, these two sub-clauses will be redone comparing the 2003 version and the IRG source glyphs.
For example, looking at what are the exact two T source glyphs for 5B0E and 5B14 as published in the charts:

T- source 5B0E	T-source 5B14
嫵	嫵
T3-4B5F	T2-565F

These represents even a better case of source separation rule than the original shown above. Clearly these two characters would have been unified if the source separation rule did not exist.

While historically T source for 5B14 had an extra stroke it has lost it in recent published versions of the standard (TCA-CNS 11643-1992 2nd plane) as shown here:

The point of Annex S.3 is to show exceptions to the source separation rules, not relate history. Concerning exception to source separation rules, the new glyphs which correspond to actual glyphs used today by the standard are a much better illustration of these exceptions.

United Kingdom: Positive with comments

Editorial comments:

E1. 4.12 code unit sequence – NOTE 2,

"Unlike previous editions of ISO/IEC 10646, this International Standard no longer uses implementation levels."
Not all previous editions use implementation levels (2nd and 3rd eds. do not).

Proposed change by UK
Reword as "Unlike some previous editions of ISO/IEC 10646, this International Standard does not use implementation levels." or similar.
Propose acceptance in principle
This can be changed to" Since its second edition: ISO/IEC 10646:2011, this International Standard does not use implementation levels."

E2. 16.6.3 Ideographic variation sequences – NOTE 2,

"This International Standard incorporates by reference the variation sequences listed in version 2010-11-14 of the Ideographic Variation Database, as described at http://www.unicode.org/ivd/data/2010-11-14/."

The version of the IVD referenced should be 2012-03-02.

Proposed change by UK
Amend to "This International Standard incorporates by reference the variation sequences listed in version 2012-03-02 of the Ideographic Variation Database, as described at http://www.unicode.org/ivd/data/2012-03-02/."
Propose acceptance

E3 – 21 Normalization forms – NOTE 4,

"Note 4" is followed by two hyphens.

Proposed change by UK
Remove the extra hyphen.
Accepted

E4 – 24.2 Name formation

"An entity names shall consist only of the following characters".
"entity names" should be singular.

Proposed change by UK
Change to "An entity name shall consist only of the following characters".
Accepted

E5 – 31.3 Character names list

"Standardized variation sequences preceded by U+2053 ~ SWUNG DASH, when this character is used as a base character in such a variation sequence."

The character given between “U+2053” and “SWUNG DASH” is U+007E ~ TILDE not U+2053 ⁓ SWUNG DASH. To avoid confusion a swung dash should be used here, as is the case in the code charts.

Proposed change by UK

Change to "Standardized variation sequences preceded by U+2053 ⁓ SWUNG DASH, when this character is used as a base character in such a variation sequence.".

Accepted

Note that in many fonts, the swung dash is very similar in appearance both in size and vertical alignment to the tilde. However a true swung dash (slightly wider) will be used in the future.

E6 – 16.6.2, 22.4, 23.2, A.4.2, Annex G, and Annex R – content linked

These sections all specify that "The content linked to is a plain text file, using ISO/IEC 646-IRV characters with LINE FEED as end of line mark".

In fact all the linked files (Allnames.txt, CJKSrc.txt, EmojiSrc.txt, HangulSy.txt, JIExt.txt, NUSI.txt, and UCSVariants.txt) are CR/LF terminated, and one file (NUSI.txt) has a BOM.

Proposed change by UK

Ensure that all linked files are terminated with LINE FEED only and do not contain a Byte Order Mark.

Accepted in principle

It is easier to change these to say CARRIAGE RETURN/LINE FEED because this is the way they are produced. The BOM will be removed from the NUSI.txt file.

USA: Negative

Technical comments:

TE.1. Annex A.5.9 288 Multilingual Latin Subset

The U.S. requests a note be added to the text of A.5.9 288 MULTILINGUAL LATIN SUBSET, clarifying that the scope of this collection is limited to that defined in ISO/IEC 9995-3, and does not cover the repertoire needed for all languages using Latin-based orthographies. The 288 MLS collection covers the letters used to write some Native American languages, for example, but does not include U+0245 LATIN CAPITAL LETTER TURNED V and U+028C LATIN SMALL LETTER TURNED V, used by Oneida and Tepehuan.

Proposed change by US:

Add a note that this collection is limited to ISO/IEC 9995-3, and does not cover all languages using Latin-based orthographies.

Propose acceptance in principle

Having no quick access to the ISO/IEC 9995-3 itself, the editor relied on the information conveyed in http://www.open-std.org/jtc1/sc35/wg1/docs/info1-9995-3.pdf to determine that indeed that the 2010 edition refers to that content. The note would say:

NOTE – The collection 288 MULTILINGUAL LATIN SUBSET does not provide an exhaustive coverage of all languages using Latin-based orthographies. It is referenced by ISO/IEC 9995-3:2010 Keyboard layouts for text and office systems -- Part 3: Complementary layouts of the alphanumeric zone of the alphanumeric section.

TE.2. Cuneiform

The U.S. requests the addition of one character to the Cuneiform block, U+12399 CUNEIFORM SIGN U U, as documented in WG2 N4493. This character was left out in error, according to a co-author of the original cuneiform proposals.

Proposed change by US:

Add the character.

Propose acceptance

The repertoire of the Cuneifrom block: 12000.123FF was revised in Amendment 2 (not yet published). It seems sensible to add it to complete the repertoire as requested by the author.

TE.3. Latin Extended-E

The U.S. requests the boundaries of the Latin Extended-E block be changed from 9 columns (U+AB30-U+ABBF) to 4 columns (U+AB30-U+AB6F). This requested change will result in a block of 5 contiguous columns, which would enable a set of 80 characters to be encoded in the BMP. This extended range would make it possible to include, for example, characters in a Cherokee Supplement block, so that the entire script could be located in the BMP. If this comment and te.4 are accommodated, we will change our vote to Yes.

Proposed change by US:

Change the end of the Latin Extended-E block from U+ABBF to U+AB6F..

WG2 discussion

The block Latin Extended-E is relatively recent (added per Amendment 1 in 2013) and only populated up to U+AB63. Furthermore, blocks are not immutable. However there is a collection 165 LATIN EXTENDED E AB30-ABBF, but it is not fixed, so in principle it is still possible to reduce its size. However should it be done, it has to be done as soon as possible.

TE.4. Siddham

The U.S. requests the 6 Siddham letter variants (U+115E0-U+115E5) be removed, since alternative approaches have been raised (such as in WG2 N4490 and N4486), and questions remain on how to best represent these letters. For these reasons, we deem these characters not mature for encoding. If this comment and te.3 are accommodated, we will change our vote to Yes.

Proposed change by US:
Remove the 6 Siddham letter variants (U+115E0-U+115E5).

WG2 discussion

Current discussions seem to indicate that the 4 letters (U+115E0-U+11E53) are not controversial. The vowel variant signs (11E54..115E5) need more discussion.

TE.5. Miscellaneous Symbols and Arrows

The U.S. requests a formal name alias be added for U+2B81 UPWARDS TRIANGLE-HEADED ARROW LEFTWARDS DOWNWARDS OF TRIANGLE-HEADED ARROW.

The name alias should be: UPWARDS TRIANGLE-HEADED ARROW LEFTWARDS OF DOWNWARDS TRIANGLE-HEADED ARROW. The formal name alias will correct an error in the names list.

Proposed change by US:
Add the formal name alias for U+2B81 as noted.

WG2 discussion

Given that the character was added very recently (Amendment 1 to the 3rd edition), we may entertain the idea of fixing the name in this edition before general availability.

Editorial comments:

E.1. Tamil

The U.S. would prefer the Lohit Tamil font be used for Tamil in the 4th edition of ISO/IEC 10646, as requested in WG2 N4476. (The font can be seen in the Tamil and Tamil Supplements blocks in PDAM1.) The new font is more aesthetically pleasing than the current font.

Proposed change by US:
Change the font.

Propose acceptance

The US is referring to Amendment 1 to the 4th edition, not Amendment 1 to the previous edition.

E.2. Latin Extended-E

The italicized comment above U+AB60 should be modified slightly from:

"These letters were used 1917-1927 in the official IPA-based Latin orthography of that era"
to:
"These letters were used from 1917 to 1927…"

Proposed change by US:
Adjust the wording as described.

Accepted

E.3. Miscellaneous Symbols and Arrows

Correct the spelling of "preferredd" in the italicized comment above U+2BEC in the names list.

Proposed change by US:
Correct the spelling as noted.
Accepted

E.4. Sharada

The first single-quote in the annotation at U+111CC ("used for writing ’matra-vowels’") should be turned the other direction.

Proposed change by US:
Fix the single quote as noted.
Accepted

E.5. Musical symbols

Remove period after annotation on U+1D1DE MUSICAL SYMBOL KIEVAL C CLEF.

Proposed change by US:
Remove the period after the annotation on U+1D1DE.
Accepted

E.6. Siddham

The italicized comment at the top of the names list, "The script is also known by the names 'Siddhamatrika' and 'Kutila'" should have the words "and ‘Kutila’" removed, since "Kutila" is too generic a term. ‘Kutila’ means ‘curled, bent’ and is found, for example, in Nepal to refer to Rañjana.

Proposed change by US:
Remove "and Kutila" from the italicized note at the top of the names list for Siddham, and change "names" to "name".
Accepted

E.7. Ahom

We recommend the font in the Ahom block be reviewed to correct clipping and winding errors.

Proposed change by US:
Review and fix the font for winding errors and clipping.
Propose acceptance in principle
There are hundreds, maybe thousands of glyphs with winding errors in the chart. The editor is relying on volunteers for these fonts. At minimum, specific on which renderings are problematic would be highly appreciated.

E.8. Early Dynastic Cuneiform

We recommend the glyphs in the Early Dynastic Cuneiform be fixed so they do not extend outside the box area for the glyphs (cf. U+12484, U+12486, U+124AB, U+12535, etc.).

Proposed change by US:
Review and fix the font so the glyphs do not extend outside the glyph boxes in the charts.
Propose acceptance in principle
These glyphs are particularly challenging for display in a grid as done in the code charts.