ISO/IEC JTC1/SC2/WG2 N 3146

Date: **2006-09-11**

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

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

Title: **Draft disposition of comments on SC2 N 3875 (PDAM text for Amendment 3.2 to ISO/IEC 10646:2003)**

Source: **Michel Suignard (project editor)**
Project: JTC1 02.10646.00.01
Status: For review by WG2
Date: 2006-04-26
Distribution: WG2
Reference: SC2 N3875, 3891
Medium: Paper, PDF file

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

A significant number of character additions not related to the amendment were made. Again, it is useful to remind all national bodies of the Action Item WG2 AI-46-10.e taken at the WG2 Xiamen meeting #46 in January 2005:

> AI-46-10.e (**All national bodies and liaison organizations**):
> To take note to restrict their ballot comments to the content of the document under ballot. All proposals for new characters or new material for the standard should be made as independent contributions outside the ballot comments. The project editor has the prerogative of ruling such proposals to be 'out of order, as not related to the text under ballot' and ignore them completely.

In consequence, such comments were ruled ‘out of scope’ and should be addressed separately. This does not preclude some of these proposed additions to be added in a future step of this amendment process.

Note – The full content of the ballot comments (minus some charts) 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***.**

Finland:Positive

Technical comments

Concerning the characters of the phonetic alphabets defined in ISO/IEC 10646 and TUS, it has become clear that some guidance for their usage would be beneficial in the standard itself.

Our concern stems from the fact that e.g. the following pairs appear confusing for the users of UPA: 1DA6 and 1D35; 1DAB and 1D38; 1DB0 and 1D3A; 1DB8 and 1D41. If a user encounters the newer phonetic character (MODIFIER LETTER SMALL CAPITAL x) before the UPA character (MODIFIER LETTER CAPITAL x), he or she is likely to make a coding error, since in practice a visual distinction between them is not clear. Although we plan to stress proper usage in the documentation for the Finnish user community, that cannot prevent possible improper usage by the international research community, thus leading to incompatible material.

We propose that appropriate wording be included in PDAM3.2 as a result of discussions between the experts at the Tokyo WG 2 meeting.

WG2 decision

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.

Technical comments

T1. Page 18 - Row 03: Greek.
With reference to ISO/IEC JTC1/SC2/WG2 N3122 "Proposal to add Latin letters and a Greek symbol to the UCS", Ireland requests that the character
 GREEK CAPITAL KAI SYMBOL be added to the PDAM at position U+03CF.
Out of scope
Document WG2 N3122 which contains a request to add sixteen Latin letters and one Greek letter (mentioned above) should be discussed separately. As a result, these characters may or may not be added in this amendment. However, the decision to add these characters in the present amendment is out of scope for this disposition of comments.

T2. Page 28 - Row 10: Myanmar.
With reference to ISO/IEC JTC1/SC2/WG2 N3115 "One additional Myanmar character for Mon for PDAM 3.2", Ireland requests that the character
 MYANMAR VOWEL SIGN E ABOVE be added to the PDAM at position U+1035.
Propose acceptance
Similar request from the US (comment T.4) and UK (comment T.2) and Myanmar is being revised by this amendment and the addition makes Mon support complete. In this context, the addition is considered 'in scope'.

T3.a Page 44 - Row 2C: Latin Extended-C - addition
With reference to ISO/IEC JTC1/SC2/WG2 N3122 "Proposal to add Latin letters and a Greek symbol to the UCS", Ireland requests that the character
 LATIN CAPITAL LETTER TURNED A be added to the PDAM at position U+ 2C6F.
Out of scope
See disposition of comment T.1.

T3.b Page 44 - Row 2C: Latin Extended-C - rename.
We also request that the name of the character U+2C78 be changed to LATIN SMALL LETTER E WITH DOUBLE FINIAL. The part of a letter known as a *finial* is usually a somewhat tapered curved end on letters such as the bottom of "C" or "e" or the top of a "double-storey a". It is this which is the distinctive feature of the U+2C78 (whose Landsmålsalfabetet name is "Stockholm e").
WG2 decision
Current name is LATIN SMALL LETTER E WITH TAIL. It would be desirable to get a reference to the proposed definition above. A finial is commonly known as an ornament decorating the upper part of another structure, such as a pinnacle or a balustrade. Besides not being a usual term in the character naming context, this would get no end of confusion for people trying to explain this name in other languages than English. It also seems to have a connotation of being the upper part of something else which does not apply here. Finally, the term tail has been used to denote similar ornaments for other Latin characters in the standard, so it does not seem useful to introduce new terminology here.

T4. Page 56 - Row A7: Latin Extended-D.
With reference to ISO/IEC JTC1/SC2/WG2 N3122 "Proposal to add Latin letters and a Greek symbol to the UCS", Ireland requests that the following characters be added to the PDAM:
 U+A779 LATIN CAPITAL LETTER INSULAR D,
 U+A77A LATIN SMALL LETTER INSULAR D,
 U+A77B LATIN CAPITAL LETTER INSULAR F,

ꝼ U+A77C LATIN SMALL LETTER INSULAR F,
ᵹ U+A77D LATIN CAPITAL LETTER INSULAR G,
ꝿ U+A77E LATIN CAPITAL LETTER TURNED INSULAR G,
ꞃ U+A77F LATIN SMALL LETTER TURNED INSULAR G,
ꞇ U+A780 LATIN CAPITAL LETTER TURNED L,
ꞇ U+A781 LATIN SMALL LETTER TURNED L,
ꞅ U+A782 LATIN CAPITAL LETTER INSULAR R,
ꞅ U+A783 LATIN SMALL LETTER INSULAR R,
ꞅ U+A784 LATIN CAPITAL LETTER INSULAR S,
ꞅ U+A785 LATIN SMALL LETTER INSULAR S,
ꞇ U+A786 LATIN CAPITAL LETTER INSULAR T, and
ꞇ U+A787 LATIN SMALL LETTER INSULAR T.

Out of scope

See disposition of comment T.1.

Editorial comments

E1. Page 18 - Row 03: Greek and Coptic.

Ireland has engaged experts in Greek typography in a review of the glyph in the ballot at U+0373 GREEK SMALL LETTER ARCHAIC SAMPI and as a result of those discussions we request that the glyph be changed to the glyph below. The capital letter is given with it for reference.

WG2 decision

It looks like that even if the new shape is adopted, it may require some refinement, is the stem on the left a hook or a new ornament? Is it as tall as the capital form?

E2. Page 28 - Row 10: Myanmar.

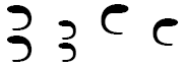
The glyph for U+1031 MYANMAR VOWEL SIGN E is incorrect. The vowel sign precedes the dotted circle.

Accepted

Note that the character was not balloted and was shown correctly in 10646:2003 .

E3. Page 56 - Row A7: Latin Extended-D.

Ireland notes that the glyphs in the ballot at U+A722..U+A725 are italic scans taken in from Gardiner's *Egyptian Grammar*. These are not the best reference glyphs for these characters. We recommend that the glyphs be changed to those shown below.



Propose acceptance

E4. Page 62 - Row A9: Rejang.

Ireland suggests that the font for Rejang be improved. It should be made somewhat bolder and the position of the diacritics with regard to the dotted circle should be clarified if possible. See the improved font on the following page. (*Not duplicated here, see SC2 3891 for reference*)

Propose acceptance

Japan, Negative with comments

Japan disapproves the document SC2 N3875 with five technical and one editorial comment. Japan will change its vote if the comments are accepted and corresponding texts are updated appropriately.

Technical comments

J1. Inclusion of CJK C1
The amendment should include appropriate change texts to add CJK Extension C1 character repertoire developed by IRG to UCS.

In "2. New tables" in "Page 30-1348 Clause 33, Code Tables and list of character names" in the amendment, add the following change texts:
Plane 02
Table *XXX* - Row *XX-XX* CJK Unified Ideographs Extension C
after
Table 210 - Row 09: Lydian
and
XXXXX-XXXXX
after
1093F
(where numbers represented by *X*'s should be assigned by WG2.)

Add a new change text in some appropriate place as follows:
Page 1351, Annex A.1
Insert the following new entry:
XXXX CJK UNIFIED IDEOGRAPHS EXTENSION C *XXXXX-XXXXX*

In page 5 (of Amd.3.2), change the following change text
382 CJK UNIFIED IDEOGRAPHS-2005 Collection 380*
9FA6-9FBB
to
382 CJK UNIFIED IDEOGRAPHS-2005 Collection 380*
9FA6-9FBB
XXX CJK UNIFIED IDEOGRAPHS-200*X* Collection 382 *
XXXXX-XXXXX

Add a new change text in some appropriate place as follows:
Page 1353, Annex A.2.3 Blocks in the SIP, after
CJK UNIFIED IDEOGRAPHS EXTENSION B 20000-2A6DF
add the following new entry:
CJK UNIFIED IDEOGRAPHS EXTENSION C *XXXXX-XXXXX*

Add detailed code charts for the new CJK Extension C1 set to be provided by IRG at the end (or somewhere appropriate) of the amendment.
Out of scope
Same rationale as for Irish comment T.1. Furthermore, it should be noted that editing instructions as provided are not necessary for addition requests. Once additions are approved by WG2, it is the task of the project editor to create the amendment instructions as appropriate. There are many more places affected by additions than the ones mentioned above.

J2 Reference to Unicode Ideographic Variation Database

In the current draft amendment, the change text for "Page 15, Sub-Clause 20.4 Variation selectors" on page 3, the allowed uses of combinations of a CJK ideographs and a Variation Selector is defined as those registered in the Ideographic Variation Database defined by Unicode Technical Standard #37. Japan is in favor of synchronizing the Unicode Ideographic Variation Database and such sequences allowed by ISO/IEC 10646, but the current amendment text contains several problems:

First problem is its dynamic nature.

UTS#37 defines database format and its registration procedure. It does not define a fixed database; the Unicode Ideographic Variation Database is assumed to grow time to time. Moreover, UTS#37 and its associated database has no way to specify one particular revision of the Ideographic Variation Database. As a result, the new sub clause 20.4 will refer to the latest revision of a dynamically changing database to consider defined versus undefined sequences of an ideograph followed by a variation selector.

It makes the conformance to the standard instable.

For example, assume you want to make your originating device to conform to 10646 with the collection 305 UNICODE 4.0 as its adopted subset. Your device needs to have a way to transmit coded representation of any legal sequence of characters as supplied by its user, and it should never transmit any illegal ones. Unless one particular ideographic variation sequence, say "CJK UNIFIED IDEOGRAPHS-4E00" followed by "VARIATION SELECTOR-99", is registered in the Unicode Ideographic Variant Database, your device should not transmit that particular sequence. Otherwise it doesn't conform to the standard. However, at the moment when that particular sequence is registered in the database, your device needs to allow its user to supply that particular sequence "CJK UNIFIED IDEOGRAPHS-4E00" followed by "VARIATION SELECTOR-99" to transmit. It seems infeasible (if not impossible) to request all conforming devices to change its behavior dynamically.

The second problem is regarding the stability of the technical specification in the indirect normative references.

The current draft amendment refers to UTS#37 as a normative reference, and the UTS#37 includes a normative reference to an external document called "Perl regular expression." specified by a URL. The maintenance procedures and/or the responsible organization of the "Perl regular express" document is unknown. Also, the wording of that particular document is informal and is different from those usually seen in the International Standards. If we add a normative reference to UTS#37 in 10646, the document called "Perl regular express" becomes a part of the International Standard. Japan considers it is not appropriate.

To solve the problems discussed above, Japan proposes the following: Each edition (including amendment) of ISO/IEC 10646 refers to one fixed revision of the Unicode Ideographic Variant Database in its normative specification as a list of USI's as its electronic attachment. The actual content of the attachment will be extracted from the original Unicode Ideographic Variant Database when revising 10646. The International Standard should also contain some informative text (as a NOTE, for example) to indicate that the date (or other information to identify the revision of the database) the contents are extracted. The NOTE can refer informatively to the UTS#37.

Thus the change text for "Page 15, Sub-Clause 20.4 Variation selectors" in the amendment three should be changed by replacing the second paragraph of the current change text as follows:

The allowed sequences using variation selectors are those defined in this clause, as well as those listed in the machine readable file of the name IVS-200*X-XX-XX*.txt distributed with this international standard. All other such sequences are undefined. Furthermore, no sequences containing variation selectors and a mix of combining characters or composite characters will be defined.

NOTE - The file IVS-200*X-XX-XX*.txt is taken from the Unicode Ideographic Variation Database as of 200*X-XX-XX*. The file contains all ideographic variation sequences defined in the database and only contains them. The Unicode Unicode Ideographic Variation Database is maintained by the Unicode Consortium with the procedure specified in Unicode Technical Standard #37, Ideographic Variation Database.

Propose acceptance in principle

Adding linked files in the standard is done in a slightly different fashion, but otherwise the request is inline with current practice for data tables within the context of ISO/IEC 10646. The alternative would be to create a link to a flat table with a version number provided by the Unicode Consortium.

J3. Definition of Extended Collection and conformance

Newly introduced notion Extended Collection and related terms have some problems in the current FPDAM text:

The current draft changes the notion of the *collection* to cover newly introduced *extended collection* and introduces a new term *regular collection* to refer to the collection in the older sense. (The term *regular collection* has no definition.) Most parts of the standard are not reworded regarding *collection*; i.e., specifications referred to regular collections in the past are changed to refer both to regular collections and extended collections silently. This type of silent change may confuse users. One way to solve this problem is to introduce the third term, e.g., *general collection* to refer both to regular collection and to extended collection, avoiding use of unqualified "collection". All texts referring to *collection* should be changed either to *general collection* or *regular collection* appropriately.

WG2 discussion
This seems overkill. The term 'regular collection' is introduced in the definition of collection itself. The same definition has a direct pointer to the extended collection. Furthermore, it is convenient having a generic 'collection' term which can be used where the distinction between 'regular' and 'extended' is not necessary. At most the term 'regular collection' could be added to the clause 4 to balance the addition of 'extended collection'.

J4. Control character names
On page 2 (of PDAM 3.2), change the names of two control characters for 000E and 000F, and insert inter-word-hyphens into some control character names in order to harmonize with the names defined in ISO/IEC 6429 as follows:

```
000E SHIFT-OUT or LOCKING-SHIFT ONE
000F SHIFT-IN or LOCKING-SHIFT ZERO
0083 NO-BREAK HERE
008E SINGLE-SHIFT TWO
008F SINGLE-SHIFT THREE
```

Note that, in ISO/IEC 6429, the actual names of some control characters are defined depending on its environment or its application. For example, those allocated 00/14 and 00/15 have different names depending on whether they are used in 7 bit environment or 8 bit environment. Including just one name is a technical error, while just put two names (as Japanese proposal above) may confuse readers who are not familiar with 6429. Simply putting names of control characters might be a bad idea. Japan is looking for a better way to satisfy the requirements from ITU-T.

Propose acceptance in principle
See also comment E.2 from UK. Note that ISO/IEC does not formally assign names as now clarified by this amendment. Strictly speaking, the ITU-T requirement is satisfied by the normative reference to ISO/IEC 6429:1992 in clause 3. All additions in the amendment are editorial notes providing non normative reference to the original text in ISO/IEC 6429. These notes can be modified to clarify that some control characters such as those coded at 000E and 000F have different ISO/IEC 6429 names depending on the environment.

J5. Use of the word CJK
On page 5 (of PDAM 3.2), the term "CJK characters" should be changed to "CJK ideographs", since the term "CJK characters" is ambiguous, and all characters contained in the referred collections are CJK ideographs.
Propose acceptance

Editorial Comments

J6. Lepcha code chart
On page 34, sample glyphs for 1C2A, 1C2B, and 1C2C look overlapping with their code point values. They need some adjustment in size.
Accepted in principle
Pending reception of the font by the contributing editor.

United Kingdom: Negative with comments:

The UK votes to DISAPPROVE the draft with the technical and editorial comments given below. If our comments are satisfactorily resolved we will change our vote to APPROVAL.

Technical comments:

T1. Page 27: Table 35 - Row 0F: Tibetan
0FD3 TIBETAN MARK INITIAL BRDA RNYING YIG MGO MDUN MA
0FD4 TIBETAN MARK CLOSING BRDA RNYING YIG MGO SGAB MA
These should be annotated "da nying yik go dun ma" and "da nying yik go kab ma" respectively, following the general practice for Tibetan character names.
Propose acceptance

T2. Page 28: Table 36 - Row 10: Myanmar
The letters added for Mon (1028, 1033..1034, 105A..1060) are not sufficient to fully represent the Mon language. In order to complete the set of letters needed for Mon, the UK requests the addition of MYANMAR VOWEL SIGN E ABOVE to Amd.3 at U+1035, as proposed in N3115.
Propose acceptance
Similar request from the US (comment T.4) and Ireland (comment T.2) and Myanmar is being revised by this amendment and the addition makes Mon support complete.

T3. Page 57 : Table 139 - Row A7: Latin Extended-D A75A LATIN CAPITAL LETTER R ROTUNDA
No evidence for this character is provided in the proposal document (N3027), and as LATIN SMALL LETTER R ROTUNDA is generally considered to be a lowercase only letter that uppercases to "R" (in the same way that LATIN SMALL LETTER LONG S uppercases to "S"), LATIN CAPITAL LETTER R ROTUNDA should be removed from Amendment 3 pending attestation of usage, and A75A left reserved.
WG2 discussion

Editorial Comments

E1. Page 1: Page 8, Sub-clause 6.3 Octet order
"The order of octet in the coded representation form" should be "The order of octets in the coded representation form".
Accepted

E2. Page 2: Page 11, Clause 15 Use of control function with the UCS
The following names do not correspond to the ISO/IEC 6429:1992 names given as aliases in the Unicode code charts. It is highly desirable that the control names given in ISO/IEC 10646 and Unicode match exactly. In particular "INDEX" is not defined in ISO/IEC 6429:1992, although it was in previous versions of the standard. The version of ISO/IEC 6429 referred to here should be made explicit.
000A LINE FEED -- Unicode has LINE FEED (LF)
000C FORM FEED – Unicode has FORM FEED (FF)
000D CARRIAGE RETURN -- Unicode has CARRIAGE RETURN (CR)
0084 INDEX -- Unicode does not provide a name (noted as "formerly known as "INDEX"")
0085 NEXT LINE -- Unicode has NEXT LINE (NEL)
Propose acceptance in principle
See also comment J4 from Japan. In fact, if you exclude the parenthetical notation, the names provided in the new note match the names described by Unicode 5.0. The latter introduced INDEX for 0084. In addition, the Unicode names do not match exactly ISO/IEC 6429 (lack of hyphen in four names as discovered by Japan). It seems advisable to fix the names as proposed by Japan and add clarification concerning 0084 INDEX that the control function INDEX was deprecated in ISO/IEC 6429:1992.

E3. Page 3 : Page 18, Clause 25 Normalization forms
"There are four normalizations forms:" should be "There are four normalization forms:"
Accepted

E4. Page 3 : Page 25, Clause 29 Named UCS Sequence Identifiers
<012F, 0307, 0301> LATIN CAPITAL LETTER I WITH OGONEK AND DOT ABOVE AND ACUTE
<012F, 0307, 0303> LATIN CAPITAL LETTER I WITH OGONEK AND DOT ABOVE AND TILDE
These should be:
<012F, 0307, 0301> LATIN SMALL LETTER I WITH OGONEK AND DOT ABOVE AND ACUTE
<012F, 0307, 0303> LATIN SMALL LETTER I WITH OGONEK AND DOT ABOVE AND TILDE
Accepted

USA: Positive with comments:

Technical comments:

T.1 Glyph correction for U+0485 and U+0486 (Cyrillic)
The US is asking for glyph correction for the following Cyrillic characters:
`U+0485 COMBINING CYRILLIC DASIA PNEUMATA` and
`U+0486 COMBINING CYRILLIC PSILI PNEUMATA`
The rationale for the correction is provided in document WG2 N3118 (L2/06-192).
Propose acceptance

T.2 Glyph correction for U+0340 and U+0341 (Combining Mark)
The US is asking for glyph correction for the following Combining Marks characters:
`U+0340 COMBINING GRAVE TONE MARK` and
`U+0341 COMBINING ACUTE TONE MARK`
Because they are canonically equivalent to U+0300 and U+0301 respectively through the normalization process, they should have exactly the same glyph as these characters in the code charts.
Propose acceptance

T.3 Name change for U+0D3D to MALAYALAM SIGN AVAGRAHA
The US is asking for a name change for the newly proposed U+0D3D from:
`U+0D3D MALAYALAM PRASLESHAM (avagraha)` to
`U+0D3D MALAYALAM SIGN AVAGRAHA (praslesham)`
Avagraha is like an apostrophe sign, and it is used in India's major scripts when Sanskrit texts are written or transliterated in them. Consider the fact that all avagraha signs encoded in major Indic scripts are encoded with character names as "X SIGN AVAGRAHA" where X is the corresponding script name as the prefix. It will be far less confusing when users compare MALAYALAM SIGN AVAGRAHA with other Indic scripts' avagraha names. The name, AVAGRAHA will bring transparency in script interoperability scenarios between Indic Scripts.
Propose acceptance

T.4 Addition of Myanmar character
The US is also supporting the addition of the following Myanmar character as proposed by document WG2 N3115 (L2/06-249):
`1035 MYANMAR VOWEL SIGN E ABOVE`
The rationale for the addition is that while this character is used in Shan, it is also used in Mon, and support for Mon is incomplete without it. In Shan, the letter represents [e] word-internally; in Mon it is used with U+102F MYANMAR VOWEL SIGN U to represent [y].
Propose acceptance
Similar request from the UK (comment T.2) and Ireland (comment T.2) and Myanmar is being revised by this amendment and the addition makes Mon support complete.

----end--