ISO/IEC JTC1/SC2/WG2 **N 3324**

Date: 2007-09-14

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 3943 (FPDAM text for Amendment 4 to ISO/IEC 10646:2003)

Source: Michel Suignard (project editor)
Project: JTC1 02.10646.04
Status: For review by WG2
Date: 2007-09-14
Distribution: WG2
Reference: SC2 N3943, N3960, WG2 N3313
Medium: Paper, PDF file

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

> Note – The full content of the ballot comments (minus some figures and 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*.

China, Negative

China is against to ISO/IEC JTC1/SC2N3943 *FPDAM4 of ISO/IEC 10646: 2003*.

(The following Chinese comments are extracted from WG2 N3313 which provided the detailed comments and are all related to Lanna.)

After carefully reading and discussing the Lanna encoding scheme in FPDAM4 and related documents, the Chinese experts in Yunnan province (where there lives over 400,00 Dai people) believe that the current scheme is uncompleted and is with errors or mistakes.

Technical comments

T1. Script name

The people which living in northern part of Thailand, Myanmar, Laos and Dai nation districts in China do not use this script as LANNA. The speakers of Northern Thai name this language "GAM MENG", means "local language", and name the script "TO MENG", means "local script", or "TO THAM", means "script of scripture". In China, people address this script "TO TAI", "TO TAILUE", "TO THAM", "script of scripture". Therefore, we regard that we should respect the people's will, using the name that people have identified.

WG2 discussion

The Chinese Member Body has made many similar requests in the past (ref WG2 N3161, N3226, and N3261). Again it is worth mentioning the ad-hoc meeting held during meeting WG2#49 (Tokyo, September 2006) which produced document WG2 N3169 which the following relevant text:

"Chinese comment I. The ad-hoc group believes that the name "Lanna" is the best name for this script in the English language. The Lanna script is used to write four languages: Khün, Lao Tham, Northern Thai, and Tai Lue. The name proposed, "Old Tai Lue" is accurate in distinction to the "New Tai Lue" script already encoded, but would not be appropriate as a generic name for this script given its use in Myanmar, Thailand and Laos."

In meeting WG2#50, this concerns was addressed by adding the notation (Old Tai Lue) to the LANNA block name and this was accepted unanimously by WG2 (resolution M50.10 and M50.13). There is no simple answer to this request because the script has many names but at the same time, the block and collection names need to be unique. Using the term 'Lanna' for the standard seems to be the best compromise, and this does not prevent member bodies to translate the 'Lanna' script name according to local preferences when referencing the standard. To resolve this issue we may also add a note to Figure 4 for Lanna, as we did for New Tai Lue.

NOTE – Lanna is also known as Meng and Tham in Thailand, or Old Tai Lue in China.

Note that the name 'Tai Lue' is used in the character names in page 9 of WG2 N3313 which is probably not acceptable.

T2. Character naming

Every alphabet has its own name in China's OLD TAILUE script project, called "TSUE TO". There are many homophone alphabets for OLD TAILUE script. Besides, different alphabets are used in different situations, and conjoin with different alphabets to express different meanings. On that account, alphabet names are for discriminating different alphabet, and making things convenient for every alphabet's use.

Most of consonant alphabets' names are their pronunciation spelled with two vowels "O" and "A", for instance, consonant k alphabet' name is "TO KO KA" and x called "TO XO XA " and so forth.

Except above alphabet name, alphabet has another name---"SING PHON", when alphabet name is used to be spelled, it is hardly to spell directly due to too many syllables. In order to make the spelling easy, every alphabet has simple "SING PHON". "SING PHON" can be equal to alphabet pronunciation, just take one syllable and spell it with vowel.

In FPDAM4 LANNA(OLD TAILUE), "SING PHON" is used for most of alphabets' names, but alphabets' names themselves are used for some cases:
1A5A LANNA SIGN MAI KANG LAI
1A5B LANNA SIGN KHUEN MAI KANG LAI
1A74 LANNA SIGN MAI KANG
1A7B LANNA SIGN MAI SAM
1AA7 LANNA SIGN MAI YAMOK

Character names in FPDAM4 are different from those in NEW TAILUE, even if the same alphabets. We believe that since "NEW TAI LUE" is already encoded in ISO/IEC 10646, character names should be conform to those encoded, especially the same characters or the same phoneme. Characters with different names in FPDAM4 and NEW TAILUE are shown below:

…(copies of New Tai Lue and Lanna name lists)

The same characters with different names are as below:

New Tai Lue	Lanna
1980	1A4E
1983	1A21
1986	1A24
1987	1A26
1988	1A27
1989	1A28
198B	1A2A
1990	1A56
1994	1A3A
199E	1A48

There are also many errors in the FPDAM4, especially the vowel alphabet names. Corrected vowel names are given below:

1A54	VOWEL SIGN AE
1A58	component part HONG
1A61	VOWEL SIGN VOWEL SHORTENER
1A65	VOWEL SIGN AM
1A66	VOWEL SIGN AM
1A67	VOWEL SIGN I
1A68	VOWEL SIGN II
1A69	VOWEL SIGN UE
1A6A	VOWEL SIGN UUE
1A6B	VOWEL SIGN U
1A6C	VOWEL SIGN UU
1A6D	FINAL NG
1A6E	VOWEL SIGN OA BELOW
1A6F	VOWEL SIGN OY
1A70	VOWEL SIGN E
1A71	VOWEL SIGN AE
1A72	VOWEL SIGN OO
1A73	VOWEL SIGN AI
1A75	VOWEL SIGN OA ABOVE
1A76	VOWEL SIGN MAI KANG
1A77	SIGN TONE-1
1A78	SIGN TONE-2

WG2 discussion
Some of the issues (name errors for the characters between 1A65-1A7B) result from a FPDAM4 production error introduced when the Vowel Sign AM and TALL AM were removed but the glyphs were not moved accordingly. See the chart in WG2 N3207R page 16 for correct representation. The rest needs ad-hoc discussion.

T3. Ordering of characters
Characters 1A4E、1A4F、1A50、1A51、1A52、1A53、1A54、1A55 are put after consonant alphabets in FPDAM4. Traditionally, these 8 glottal alphabets are put preceding all consonant alphabets, according to consonant's articulate segment attribution collate: glottal、velar、apical、labial. Thus, these 8 glottal should be ordered preceding all alphabets.

In FPDAM4, the same phoneme's characters are not arranged altogether. For instance, 6 phoneme "S" characters 1A28、1A29、1A49、1A4A、1A4B、1A5F are located at three positions respectively.
Propose non acceptance
Typically, order of character in character encoding is non important because all ordering schemes require their own weight tables. Even the ASCII subset of the Latin script is not encoded according to any common

ordering rules. Furthermore the encoding sequence of the Lanna script follows common encoding sequencing used for other similar scripts such as New Tai Lue. Finally, given the experience of the production error, it seems unwise to re-order the script at this phase of the amendment process.

T4. Characters should be deleted (1A28, 1A55, 1A5D, 1A72)
Some characters in FPDAM4 are the same characters in different handwriting, they are unlikely to appear in the same version, so it is unnecessary to be encoded. For example:
  1A28 v.s. 1A4B: 1A28 is the different writing form of 1A4B, these two characters pronunciation and usage are totally the same, thus the encoding of **1A28** is not needed.
  1A71 v.s. 12A7: the same as that of 1A28 v.s. 1A4B, thus **1A72** is not needed.

1A58 is a combining mark , it can combined with k、t、K、T、U、h、s、z、a . In FPDAM, **1A55** is a combination of 1A58 and 1A4E , so 1A55 is unnecessary to be encoded .

The combined character **1A5D** has only one combining form in practice, i.e. 1A50. There is no need to encode 1A5D any more since 1A50 is encoded separately.
<u>**WG2 discussion**</u>
This requires ad-hoc discussion. Note that because of the erroneous shifting in 1A65-1A7B, it is not clear whether the comment about 1A71 and 1A72 reflects the name or the glyph.

T5.a. Characters should be added (1A7C, 1A7D)
The frequently used super joined characters 1A7C and 1A7D, which is proposed in N3207, are not in FPDAM4. 1A7C and 1A7D See characters table N3207:

(Copy of page 15 of original WG2 N3207, superseded by page 14 of WG2 N3207R which shows these characters in position 1A7A and 1A7B)

<u>**Accepted**</u>
The name list already contains them. This is again the result of the production error mentioned earlier. The glyphs will be corrected to show that these two characters are indeed encoded.

T5.b. Characters should be added (1A5E, 1A60, 1A61)
There three characters need to be encoded: 1A5E、 1A60 1A61, see character table Old Tai Lue. The Characters are named "TO TAI", means "dead character", cannot be separated.

Our new character table:..
<u>**WG2 discussion**</u>
Note that because of the re-ordering this should read:
1A5E TAI LUE LETTER FINAL M
1A60 TAI LUE LETER FINAL K
1A61TAI LUE LETTER FINAL K
This requires ad-hoc discussion. Note that because of the erroneous shifting in 1A65-1A7B, it is not clear whether the comment about 1A71 and 1A72 reflects the name or the glyph. Finally many character names in this proposal are duplicated, which is not acceptable.

<u>Germany, Positive</u>

Editorial comments

E1. Capital Letter Sharp S glyph change
LATIN CAPITAL LETTER SHARP S We request a glyph change for U+1E9E in the code chart.
<u>**Accepted**</u>
It is assumed that the glyph change is the same as sought by Irish comment E.2

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 20, Table 36 - Row 10: Myanmar.
With reference to ISO/IEC JTC1/SC2/WG2 N3277R "Proposal for encoding additional Myanmar characters for Shan in the UCS", Ireland requests the following changes to the Myanmar characters which are under ballot:
Move U+1077 MYANMAR LETTER SHAN CA to U+1078
Move U+1078 MYANMAR LETTER SHAN NYA to U+107A and change its glyph as in N3277R.
Move U+1079 MYANMAR LETTER SHAN NA to U+107C
Move U+107A MYANMAR LETTER SHAN PHA to U+107D
Move U+107B MYANMAR LETTER SHAN FA to U+107E
Move U+107C MYANMAR LETTER SHAN THA to U+1080
Move U+107D MYANMAR LETTER SHAN HA to U+1081
Move U+107E MYANMAR CONSONANT SIGN SHAN MEDIAL WA to U+1082
Move U+107F MYANMAR VOWEL SIGN SHAN AA to U+1083
Move U+1080 MYANMAR VOWEL SIGN SHAN E to U+1084
Move U+1081 MYANMAR VOWEL SIGN SHAN E ABOVE to U+1085
Move U+1082 MYANMAR VOWEL SIGN SHAN FINAL Y to U+1086
Move U+1083 MYANMAR SIGN SHAN TONE-2 to U+1087
Move U+1084 MYANMAR SIGN SHAN TONE-3 to U+1088
Move U+1085 MYANMAR SIGN SHAN COUNCIL TONE-4 to U+108C and rename it MYANMAR SIGN SHAN COUNCIL TONE-3
Move U+1086 MYANMAR SIGN SHAN TONE-5 to U+1089
Move U+1087 MYANMAR SIGN SHAN TONE-6 to U+108A
Move U+1088 MYANMAR SIGN SHAN COUNCIL EMPHATIC to U+108D
Move U+1089 MYANMAR LETTER RUMAI PALAUNG FA to U+108E
Move U+108A MYANMAR SIGN RUMAI PALAUNG TONE-5 to U+108F
Add MYANMAR LETTER SHAN GA to U+1077
Add MYANMAR LETTER SHAN ZA to U+1079
Add MYANMAR LETTER SHAN DA to U+107B
Add MYANMAR LETTER SHAN BA to U+107F
Add MYANMAR SIGN SHAN COUNCIL TONE-2 to U+108B
Add MYANMAR SHAN DIGIT ZERO to U+1090
Add MYANMAR SHAN DIGIT ONE to U+1091
Add MYANMAR SHAN DIGIT TWO to U+1092
Add MYANMAR SHAN DIGIT THREE to U+1093
Add MYANMAR SHAN DIGIT FOUR to U+1094
Add MYANMAR SHAN DIGIT FIVE to U+1095
Add MYANMAR SHAN DIGIT SIX to U+1096
Add MYANMAR SHAN DIGIT SEVEN to U+1097
Add MYANMAR SHAN DIGIT EIGHT to U+1098
Add MYANMAR SHAN DIGIT NINE to U+1099
Add MYANMAR SYMBOL SHAN ONE to U+109E
Add MYANMAR SYMBOL SHAN EXCLAMATION to U+109F.
Propose acceptance
The move and renaming are also requested by US comment T.4. The additions request is similar to what is proposed in document WG2 N3321 concerning the 17 Shan characters.

T2. Page 42, Table 111 - Row 2E: Supplemental Punctuation.
The arrangement of the code table is unsatisfactory. We request that the characters in columns 2E2, 2E3, and 2E4 be rearranged as shown in the chart on page 5 below. We also request a name change to three of the characters, whose names were proposed in N3193 as TILDE but which were proposed to be changed by the UTC to SWUNG DASH. The TILDE U+007F and the SWUNG DASH U+2053 are similar characters with similar functions. Both may be used in lexicography as a sign of repetition (Figs. 1, 2, 3 below). In Uralic linguistics (as noted in N2419), the SWUNG DASH is used to indicate relationship between the entities on either side of it. This relation can be etymological (Figs. 4, 5 below) or phonological/phonetic (Figs. 6, 7, 8 below).

(Examples following)

The SWUNG DASH tends to be half again as long as the TILDE, however, and sometimes has fancy curved serifs. In German lexicography, the SWUNG DASH is unknown, and indeed, as shown below, the term used in German is "Tilde". Figures 1 and 2 above show quite clearly that the German character is much smaller than the typical SWUNG DASH.

Compare also the following character and their glyphs, taken from Helvetica and Lucida Grande, two fonts which ship with the Mac OS:

(screen dumps)

The characters shown are U+007F TILDE, U+00E3 LATIN SMALL LETTER A WITH TILDE, U+02DC SMALL TILDE, and U+2053 SWUNG DASH. The three German lexicographical characters under ballot are short TILDEs, not long SWUNG DASHes. We have seen no evidence that SWUNG DASH (with or without dots or rings) is used in German lexicography. The names should be changed.

WG2 discussion

We are assuming that the current arrangement was made to leave room for a forthcoming proposal. If none are done by this meeting, the re-arrangement should be OK. The name change needs to be discussed. The fact that the switch to SWUNG DASH was requested by the UTC is not really relevant to this discussion. The current names (with SWUNG DASH) were in WG2 N3211, referenced by resolution M50.26 adding these characters, and approved unanimously at the WG2 meeting #50. Note that glyph for 2E19 PALM BRANCH is different from the original glyph (intended?).

Editorial comments

E1. Page 18, Table 30 - Row 0D: Malayalam.
The character U+0D7C MALAYALAM LETTER CHILLU RR is a form that represents both U+0D30 MALAYALAM LETTER RA and U+0D31 MALAYALAM LETTER RRA, and the character U+0D7D MALAYALAM LETTER CHILLU L is a form that represents both U+0D24 MALAYALAM LETTER TA and U+0D32 MALAYALAM LETTER LA (and sometimes U+0D26 MALAYALAM LETTER DA). The confusion has to do with the fact that in pre-pausal position the sounds fell together. We propose that the characters be annotated as follows:

    MALAYALAM LETTER CILLU RR
      = historically derived from the full letter ra
      = often used for chillu r
    MALAYALAM LETTER CILLU L
      = historically derived from the full letter ta
      = used for chillu t and sometimes for chillu

Propose acceptance in principle

These annotations (with correction as below) could be added in the new format proposed by Amendment 5, but can be already captured in the Unicode name list format. Note that the character names contain 'CHILLU', not 'CILLU'.

E2. Page 24, Table 71 - Row 1E: Latin Extended Additional.
Ireland requests that the glyph of U+1E9E LATIN CAPITAL LETTER SHARP S be changed from the glyph on the left to the glyph on the right.

Propose acceptance

Similar to comment E.1 from Germany.

E3. Row 03: Combining Diacritical Marks. Ireland has been made aware of a glyph error in the code charts for Unicode Version 5.0, the glyphs for U+0333 and U+0347 are incorrect. The following should be used:

Propose acceptance

Japan, Negative

Japan votes against SC2 N3943 (ISO/IEC 10646:2003 FPDAM4) with the following comments. Japan will change its vote if they are accepted accordingly.

Technical comments
Regarding the addition of a new character CJK UNIFIED IDEOGRAPH-9FC3, the following two points are inappropriate and should be fixed:
- (a) The G source reference "G_KX0809" looks strange. The number seems to represent a page of Kangxi dictionary, but the page shows both 93FC and 4039 (where the new character was considered to be unified before). It is not helpful for identification of the character to say "on page 0809". It should say something like "G_KX0809.020". If the detailed identification is not needed here, it should simply say "G_KX".
- (b) Because the KP source for the new 9FC3 is moved from a compatibility ideograph FAD4, the file CJKC_SR.txt should be updated as well as CJKU_SR.txt, with the following instruction: Remove the KP1-5E2B source from 0FAD4 entry in CJKC_SR.txt

Propose acceptance in principle

a) For Amendment 4 it should probably just say G_KX, because Amendment 4 does not change the CJK source file format while Amendment 5 does, so given the current constraint there was no solution to fully convey resolution M50.20. However Amendment 5 should then contain the full reference for that new character. Note that currently Amendment 5 has a G_KXddddd format which does not allow the referencing shown above. The UK comment T.2 addresses the same issue and propose to use G_KXdddddd (note 6 'd' instead of 5). See Amendment 5 (PDAM5) disposition of comments for further details, especially the Japanese section.
b) The dereferencing of the KP1 source in CJKC_SR.txt makes sense.

UK, Positive with comments

The UK votes to APPROVE the amendment, with the following technical comments.

Technical comments

T.1 Page 2 : Source references for CJK Unified Ideographs
"Remove the T4-3946 source from 04039 entry;"
In addition to removing the T4-3946 source reference from 04039, it was agreed in Resolution M50.20 to add the T6-4B7A source reference (originally for 2F949) to 04039. Therefore the above line should be changed to something like "Replace the T4-3946 source in the 04039 entry with T6-4B7A;" and the entry for 04039 in CJKU_SR.txt changed to
04039;G3-5952;T6-4B7A;J4-7222;;;H-98E6;KP1-5E34;;
Propose acceptance
Also part of US comment T.1

T.2 Page 3 : G Source Reference for 09FC3
The new G-source reference for 09FC3 is given as "G_KX0809" in the multi-column chart on page 3, but in CJKU_SR.txt it is given as "G_KX00809", and in Resolution M50.20 it is "G_KX0809.020". In fact, on the basis of the format of the G_KX source references in Amd.5 (G_KXppppnn, where pppp is the page number and nn is the position of the character on this page), we believe that the source reference should be "G_KX080902" (i.e. second character on page 809 of the Kangxi Dictionary). The multi-column chart on page 3 and the contents of CJKU_SR.txt should both be amended to specify "G_KX080902".

In addition, as this is a new source reference format, the description of "2nd field: Hanzi G sources" in Sub-clause 27.1 (Source references for CJK Unified Ideographs) should be amended to add "(G_KXdddddd)" after "(G_KX)".
Propose accepted in principle
See disposition of Japanese comments

USA: Positive with comments:

The US National Body is voting Yes with comments on the following SC2 ballot.

Technical Comments:

T.1 Source reference for existing CJK characters (page 2)

The WG2 Resolution M50.20 (dis-unification of CJK Unified ideograph 4039) was not completely conveyed in amendment 4. In addition to the removal of T4-3946, it also needs to be replaced by T6-4B7A. Furthermore, the linked file CJKU_SR.txt has none of these changes for the U+04039 entry. Note that the U+09FC3 is correct, and that this also needs to be fixed in the proposed amendment 5 which contains the CJKU_SR.txt file.

Propose acceptance

T.2 New collection numbers (page 3)

In page 3 of the amendment, the collections number for the proposed scripts are incorrect because they re-use number already allocated to scripts added by amendment 3. So the text after '112 ARABIC SUPPLEMENT' should read:

<<

In the list of collection numbers and names, after

139 REJANG

insert new entries as follows:

140 LANNA	1A20-1AAF
141 CYRILLIC EXTENDED-A	2DE0-2DFF *
142 CYRILLIC EXTENDED-B	A640-A69F
143 CHAM	AA00-AA5F

>>

These collection numbers are again used in the next page (ref Note 3 about keywords), and therefore they should be updated there as well.

Propose acceptance

T.3 Deprecation of existing collections (page 3)

The collection 271 COMBINING CHARACTERS B-2 is specified by direct referencing annex B.2 that this amendment is removing. Therefore it needs to be deprecated with wording similar to other deprecated collections (such as collections 400 and 500). The text for the definition of collection 270 COMBINING CHARACTERS and 1900 SMP COMBINING CHARACTERS also needs to be changed to point to annex B instead of annex B.1.

Propose acceptance

T.4 Shan characters (Myanmar)

The US is in favor of the re-ordering of the Shan characters as shows in WG2 N3277 which are already in the amendment (these are shown in blue in the document) to make room for additional characters (shown in yellow) in the future. N3277 also proposes a glyph change for 107A (formerly 1078) and name change for 108C (formerly 1085) from 'MYANMAR SIGN SHAN COUNCIL TONE-4' to 'MYANMAR SIGN SHAN COUNCIL TONE-3' which are agreeable by the US. In addition, the glyph for 1040 MYANMAR DIGIT ZERO should reflect actual use. If the glyph is identical to the one used for 101D MYANMAR LETTER WA, there should not be an artificial difference as currently shown in N3277.

Propose acceptance

Similar comments to Irish T.1 (minus the additions covered by document WG2 N3321 concerning the 17 Shan characters). The US is also asking for a glyph change for 1040 MYANMAR DIGIT ZERO.

T.5 Malayalam Chillu characters

The US NB does not support a name change for 0D7C MALAYALAM LETTER CHILLU RR.

Noted

T.6 Latin Epigraphic characters

The US NB does not support the addition of lower case forms to these characters.

Noted

T.7 Mirrored characters

After further study the US NB is in favor of undoing the change made by Amendment 3 concerning the mirrored property of the following characters.

```
2018    LEFT SINGLE QUOTATION MARK
2019    RIGHT SINGLE QUOTATION MARK
201A    SINGLE LOW-9 QUOTATION MARK
201B    SINGLE HIGH-REVERSED-9 QUOTATION MARK
201C    LEFT DOUBLE QUOTATION MARK
201D    RIGHT DOUBLE QUOTATION MARK
201E    DOUBLE LOW-9 QUOTATION MARK
201F    DOUBLE HIGH-REVERSED-9 QUOTATION MARK
301D    REVERSED DOUBLE PRIME QUOTATION MARK
301E    DOUBLE PRIME QUOTATION MARK
301F    LOW DOUBLE PRIME QUOTATION MARK
```

As a result these characters should not be mirrored in bidirectional context.

Propose acceptance

Editorial Comments:

E.1 Definition of UNICODE 5.1 in A.6.6 (page 4)

Correct the definition as follows:

308   The fixed collection UNICODE 5.1 is arranged by planes as follows.

Propose acceptance

E.2 Identification of UTF-16 (page 7)

In the updated paragraph replace 'sequence and note' with 'sequence'

Propose acceptance

E.3 Identification of UTF-8 (page 7)

In the updated paragraph replace 'by following' with 'by the following'

Propose acceptance

E.4 Identification ASN.1 character abstract syntaxes (page 7)

Fix the Note 4 and following paragraph as follows:

> NOTE 4 – As an example, the object identifier for the subset comprising the collections BASIC LATIN, LATIN-1 SUPPLEMENT, and MATHEMATICAL OPERATORS is:
>
> {iso standard 10646 (0) level-3 (3) collections (1) 1 2 39}

ISO/IEC 8824 also specifies object descriptors corresponding to object identifier values. For an unrestricted repertoire, the corresponding object descriptor is as follows:

Propose acceptance