ISO/IEC JTC 1/SC 2/WG 2 **N2380R**

DATE: 2001-10-14

ISO/IEC JTC 1/SC 2/WG 2
Universal Multiple-Octet Coded Character Set
(UCS)

DOC TYPE:	National Body Contribution
TITLE:	Cambodian official objection to the existing Khmer block in UCS
SOURCE:	Committee for Standardization of Khmer Characters in Computers*
PROJECT:	JTC 1.02.18 – ISO/IEC 10646
STATUS:	For discussion at the 41st WG2 Meeting in Singapore
ACTION ID:	ACT
DISTRIBUTION:	ISO/IEC JTC 1/SC 2/WG 2
MEDIUM:	Electronic
NO. OF PAGES:	9

* Standard Organization accredited by Industrial Standards Bureau of Cambodia (ISC)

Cambodian official objection to the existing Khmer block in UCS

We are pleased to have this opportunity to present the official views of Cambodia to the relevant Working Group and Sub-Committee of JTC1.

The Committee for Standardization of Khmer Characters in Computers seeks a rescission of the Khmer Code Table as published in ISO/IEC 10646-1 2nd edition, 2000, and its complete replacement by the character set being prepared as a Cambodian national standard (see Appendix One).

We base our request on the following grounds:

1) no appropriate official Cambodian representative participated in any of the discussions leading to the adoption of the current code table by ISO/IEC, and this code table has never been officially endorsed within Cambodia;

2) the present code table contains major deficiencies as outlined in Appendix Two, of which the most significant is the decision not to allocate individual code points for the subscript consonants, but instead to follow the "virama model", presenting severe inconsistency in the light of Khmer orthography and causing unnecessary inefficiency for Khmer character processing, which would not be faced by the replacement table. We are also against the further attempt to impose the virama model as proposed in N2359.

We realize that this is an unusual request in the light of the stated position of ISO/IEC 10646-1 that "the names and allocation of the characters … will remain unchanged" (p.9). However, it is our contention that as due process was not followed in the adoption of the Khmer code table it is entirely within the norms of international standard-setting for the issue to be revisited now that a request has been formally submitted by the appropriate national body.
Furthermore, we understand that a complete replacement of a published code table within ISO/IEC 10646-1 is not without precedent.

We look forward to the opportunity for the Cambodian delegation to discuss this issue at the WG2 and SC2 meetings in Singapore, 15-19 October 2001.

Dr Pan Sorasak,
Under Secretary of State of the Council of Ministers, Royal Government of Cambodia
Deputy Chairman of the Committee for Standardization of Khmer Characters in Computers

Phnom Penh, 8 October 2001

Appendix One

Cambodian Standard Coded Character Set (CSCCS)

1. Scope

This Cambodian standard specifies Cambodian Standard Coded Character Set (CSCCS) as a basis for computerizing Khmer script.

CSCCS defines two things:

1) A character set that contains all the necessary Khmer characters that can be seen in modern documents. The basic source is so-called Chuon Nath's dictionary ("Dictionaire Cambodgien", 5e édition, Institut Boudhique, 1967-1968), the well-known standard dictionary for the modern Khmer script in Cambodia.
2) The relative code positions of these characters. The positions of characters are determined so that binary sorting can produce as good a result as possible in light of Chuon Nath's dictionary.

CSCCS does not define the absolute code value of each character that will be used in a concrete device. The concrete encoding schemes will be defined in another Cambodian standard.

CSCCS is based on the distinction between a character and a glyph. CSCCS gives a code value only to each character, premising that a rendering device will produce proper glyphs to represent characters. However, CSCCS does not preclude another Cambodian standard from encoding each glyph instead of each character, as long as any glyph sequence can be unambiguously converted to the corresponding character sequence. CSCCS does not define the relationship between a particular character sequence and a particular glyph sequence. This will be done in another Cambodian standard.

2. References

In order to secure coexistence with other scripts, the frameworks of the following international standards are suggestive:

ISO/IEC 10646-1:2000, Information technology — Universal Multiple-Octet Coded Character Set (UCS) — Architecture and Basic Multilingual Plane. * Except the existing "Khmer" block

ISO/IEC 2022:1994, Information technology — Character code structure and extension techniques.

3. Terminologies

Binary Sorting: Ordering characters according to their code values.
Character: A unit of information necessary to organize, control or represent textual data.
Code: A system of numerical expression for characters.
Coded Character Set: A set of characters each of which has a code value.
Code Position: A position given to a character or glyph in a coded character set.
Code Value: A concrete numerical expression given to a character.
Device: A hardware or software component for information processing.
Encoding: To give a code value to each character.
Glyph: A unit of graphical expression of a character or characters. One character may have several different glyphs according to context. One character may be composed of multiple glyphs, while one glyph may represent multiple characters.
Rendering: To produce a proper glyph sequence from a character sequence according to context.

4. The Coded Character Set

4.1 Code Table

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	ក	ថ	អ	◌៉	◌ួ	◌្ក	◌្ថ	◌្អ	០	០ ៨	៨ ៨	○				
1	ខ	ទ	ឥ	◌៊	◌ើ	◌្ខ	◌្ទ		១	១ ។	។ ១	∧				
2	គ	ធ	ឦ	◌៌	◌ឿ	◌្គ	◌្ធ		២	២ ។	។ ២	ǀ				
3	ឃ	ន	ឧ	◌៍	◌ៀ	◌្ឃ	◌្ន	◌្ឧ	៣	៣ ។	។ ៣	∧∧				
4	ង	ប	ឩ	◌់	◌េ	◌្ង	◌្ប		៤	៤ ។	។ ៤	∨				
5	ច	ផ	ឪ	◌ះ	◌ែ	◌្ច	◌្ផ		៥	៥ ។	។ ៥	ɤ				
6	ឆ	ព	ឫ	◌័	◌ៃ	◌្ឆ	◌្ព	◌្ឫ	៦	៦ ។	។ ៦	∖				
7	ជ	ភ	ឬ	◌៑	◌ោ	◌្ជ	◌្ភ		៧	៧ ។	។ ៧	ᴎ				
8	ឈ	ម	ឭ	◌ិ	◌ៅ	◌្ឈ	◌្ម		៨	៨ ។	។ ៨	/				
9	ញ	យ	ឮ	◌ា	◌ុំ	◌្ញ	◌្យ		៩	៩ ។	។ ៩	ƨ				
A	ដ	រ	ឯ	◌ី	◌ំ	◌្ដ	◌្រ	◌្ឯ	៛	១០ ។	។ ១០	S				
B	ឋ	ល	ឰ	◌ឹ	◌ាំ	◌្ឋ	◌្ល		៖	១១ ។	។ ១១	V1				
C	ឌ	វ	ឱ	◌ឺ	◌ៈ	◌្ឌ	◌្វ		។	១២ ។	។ ១២					
D	ឍ	ស	ឲ	◌ឹ	◌៎	◌្ឍ	◌្ស		៕	១៣ ។	។ ១៣					
E	ណ	ហ	ឝ	◌ុ	◌៏	◌្ណ	◌្ហ	◌្ឝ	៙	១៤ ។	។ ១៤					
F	ត	ឡ	ឞ	◌ូ	ៗ	◌្ត		◌្ឞ	៚	១៥ ។	។ ១៥					

4.2 Character name

Consonants

00	ក	KHMER CONSONANT KA
01	ខ	KHMER CONSONANT KHA
02	គ	KHMER CONSONANT KO
03	ឃ	KHMER CONSONANT KHO
04	ង	KHMER CONSONANT NGO
05	ច	KHMER CONSONANT CA
06	ឆ	KHMER CONSONANT CHA
07	ជ	KHMER CONSONANT CO
08	ឈ	KHMER CONSONANT CHO
09	ញ	KHMER CONSONANT NHO
0A	ដ	KHMER CONSONANT DA
0B	ឋ	KHMER CONSONANT TTHA
0C	ឌ	KHMER CONSONANT DO
0D	ឍ	KHMER CONSONANT TTHO
0E	ណ	KHMER CONSONANT NA
0F	ត	KHMER CONSONANT TA
10	ថ	KHMER CONSONANT THA
11	ទ	KHMER CONSONANT TO
12	ធ	KHMER CONSONANT THO
13	ន	KHMER CONSONANT NO
14	ប	KHMER CONSONANT BA
15	ផ	KHMER CONSONANT PHA
16	ព	KHMER CONSONANT PO
17	ភ	KHMER CONSONANT PHO
18	ម	KHMER CONSONANT MO
19	យ	KHMER CONSONANT YO
1A	រ	KHMER CONSONANT RO
1B	ល	KHMER CONSONANT LO
1C	វ	KHMER CONSONANT VO
1D	ស	KHMER CONSONANT SA
1E	ហ	KHMER CONSONANT HA
1F	ឡ	KHMER CONSONANT LA
20	អ	KHMER CONSONANT QA

Independent vowels

21	ឥ	KHMER INDEPENDENT VOWEL QI
22	ឦ	KHMER INDEPENDENT VOWEL QII
23	ឧ	KHMER INDEPENDENT VOWEL QU
24	ឩ	KHMER INDEPENDENT VOWEL QUU
25	ឪ	KHMER INDEPENDENT VOWEL QUUV
26	ឫ	KHMER INDEPENDENT VOWEL RY
27	ឬ	KHMER INDEPENDENT VOWEL RYY
28	ឭ	KHMER INDEPENDENT VOWEL LY
29	ឮ	KHMER INDEPENDENT VOWEL LYY
2A	ឯ	KHMER INDEPENDENT VOWEL QE
2B	ឰ	KHMER INDEPENDENT VOWEL QAI
2C	ឱ	KHMER INDEPENDENT VOWEL QOO
2D	ឳ	KHMER INDEPENDENT VOWEL QAU

Pali/Sanskrit extending consonants

2E	ឝ	KHMER CONSONANT SHA
2F	ឞ	KHMER CONSONANT SSA

Diacritic signs

30	◌៍	KHMER SIGN TOANDAKHEAT
31	◌៏	KHMER SIGN AHSDA
32	◌៌	KHMER SIGN ROBAT
33	◌៎	KHMER SIGN KAKABAT
34	◌់	KHMER SIGN BANTAK
35	◌ះ	KHMER SIGN YUKALEAKPINTU
36	◌័	KHMER SIGN SAMYOKSANNHA
37	◌៑	KHMER SIGN VIREAM
38	◌៝	KHMER SIGN ATTHACAN

Dependent vowel signs

39	◌ា	KHMER VOWEL SIGN SRAK AA
3A	◌ិ	KHMER VOWEL SIGN SRAK I
3B	◌ី	KHMER VOWEL SIGN SRAK II
3C	◌ឹ	KHMER VOWEL SIGN SRAK Y
3D	◌ឺ	KHMER VOWEL SIGN SRAK YY
3E	◌ុ	KHMER VOWEL SIGN SRAK U
3F	◌ូ	KHMER VOWEL SIGN SRAK UU
40	◌ួ	KHMER VOWEL SIGN SRAK UA
41	◌ើ	KHMER VOWEL SIGN SRAK OE
42	◌ឿ	KHMER VOWEL SIGN SRAK YA
43	◌ៀ	KHMER VOWEL SIGN SRAK IE
44	◌េ	KHMER VOWEL SIGN SRAK E
45	◌ែ	KHMER VOWEL SIGN SRAK AE
46	◌ៃ	KHMER VOWEL SIGN SRAK AI
47	◌ោ	KHMER VOWEL SIGN SRAK OO
48	◌ៅ	KHMER VOWEL SIGN SRAK AU

49	◌ុំ	KHMER VOWEL SIGN SRAK OM
4A	◌ំ	KHMER VOWEL SIGN SRAK AM
4B	◌ាំ	KHMER VOWEL SIGN SRAK AAM
4C	◌ះ	KHMER VOWEL SIGN SRAK AH

Consonant shifter signs

4D	◌៉	KHMER SIGN MUSEKATOAN
4E	◌៊	KHMER SIGN TREISAP

Repeater sign

4F	ៗ	KHMER SIGN LEKTO

Subscript consonant signs

50	◌្ក	KHMER CONSONANT SIGN COENG KA
51	◌្ខ	KHMER CONSONANT SIGN COENG KHA
52	◌្គ	KHMER CONSONANT SIGN COENG KO
53	◌្ឃ	KHMER CONSONANT SIGN COENG KHO
54	◌្ង	KHMER CONSONANT SIGN COENG NGO
55	◌្ច	KHMER CONSONANT SIGN COENG CA
56	◌្ឆ	KHMER CONSONANT SIGN COENG CHA
57	◌្ជ	KHMER CONSONANT SIGN COENG CO
58	◌្ឈ	KHMER CONSONANT SIGN COENG CHO
59	◌្ញ	KHMER CONSONANT SIGN COENG NHO
5A	◌្ដ	KHMER CONSONANT SIGN COENG DA
5B	◌្ឋ	KHMER CONSONANT SIGN COENG TTHA
5C	◌្ឌ	KHMER CONSONANT SIGN COENG DO
5D	◌្ឍ	KHMER CONSONANT SIGN COENG TTHO
5E	◌្ណ	KHMER CONSONANT SIGN COENG NA
5F	◌្ត	KHMER CONSONANT SIGN COENG TA
60	◌្ថ	KHMER CONSONANT SIGN COENG THA
61	◌្ទ	KHMER CONSONANT SIGN COENG TO
62	◌្ធ	KHMER CONSONANT SIGN COENG THO
63	◌្ន	KHMER CONSONANT SIGN COENG NO
64	◌្ប	KHMER CONSONANT SIGN COENG BA
65	◌្ផ	KHMER CONSONANT SIGN COENG PHA
66	◌្ព	KHMER CONSONANT SIGN COENG PO
67	◌្ភ	KHMER CONSONANT SIGN COENG PHO
68	◌្ម	KHMER CONSONANT SIGN COENG MO
69	◌្យ	KHMER CONSONANT SIGN COENG YO
6A	◌្រ	KHMER CONSONANT SIGN COENG RO
6B	◌្ល	KHMER CONSONANT SIGN COENG LO
6C	◌្វ	KHMER CONSONANT SIGN COENG VO
6D	◌្ស	KHMER CONSONANT SIGN COENG SA
6E	◌្ហ	KHMER CONSONANT SIGN COENG HA
6F	▨	<reserved>
70	◌្អ	KHMER CONSONANT SIGN COENG QA

Subscript independent vowel signs

71	▨	<reserved>
72	▨	<reserved>
73	◌្ឧ	KHMER VOWEL SIGN COENG QU
74	▨	<reserved>
75	▨	<reserved>
76	◌្ឫ	KHMER VOWEL SIGN COENG RY
77	▨	<reserved>
78	▨	<reserved>
79	▨	<reserved>
7A	◌្ឯ	KHMER VOWEL SIGN COENG QE
7B	▨	<reserved>
7C	▨	<reserved>
7D	▨	<reserved>

Pali/Sanskrit extending subscript consonant signs

7E	◌្ឝ	KHMER CONSONANT SIGN COENG SHA
7F	◌្ឞ	KHMER CONSONANT SIGN COENG SSA

Digits

80	០	KHMER DIGIT ZERO
81	១	KHMER DIGIT ONE
82	២	KHMER DIGIT TWO
83	៣	KHMER DIGIT THREE
84	៤	KHMER DIGIT FOUR
85	៥	KHMER DIGIT FIVE
86	៦	KHMER DIGIT SIX
87	៧	KHMER DIGIT SEVEN
88	៨	KHMER DIGIT EIGHT
89	៩	KHMER DIGIT NINE

Currency symbol

8A	៛	KHMER CURRENCY SYMBOL RIEL

Punctuation signs

8B	៖	KHMER SIGN CAMNOCPIIKUH
8C	។	KHMER SIGN KHAN
8D	៕	KHMER SIGN BARIYOSAN
8E	៙	KHMER SIGN PHNEKMOAN
8F	៚	KHMER SIGN KOMOT

Lunar date symbols

90	᧠	KHMER SYMBOL PATHAMASAT
91	᧡	KHMER SYMBOL MUOY KOET
92	᧢	KHMER SYMBOL PII KOET
93	᧣	KHMER SYMBOL BEI KOET
94	᧤	KHMER SYMBOL BUON KOET
95	᧥	KHMER SYMBOL PRAM KOET
96	᧦	KHMER SYMBOL PRAM-MUOY KOET
97	᧧	KHMER SYMBOL PRAM-PII KOET
98	᧨	KHMER SYMBOL PRAM-BEI KOET
99	᧩	KHMER SYMBOL PRAM-BUON KOET
9A	᧪	KHMER SYMBOL DAP KOET
9B	᧫	KHMER SYMBOL DAP-MUOY KOET
9C	᧬	KHMER SYMBOL DAP-PII KOET
9D	᧭	KHMER SYMBOL DAP-BEI KOET
9E	᧮	KHMER SYMBOL DAP-BUON KOET
9F	᧯	KHMER SYMBOL DAP-PRAM KOET
A0	᧰	KHMER SYMBOL TUTEYASAT
A1	᧱	KHMER SYMBOL MUOY ROC
A2	᧲	KHMER SYMBOL PII ROC
A3	᧳	KHMER SYMBOL BEI ROC
A4	᧴	KHMER SYMBOL BUON ROC
A5	᧵	KHMER SYMBOL PRAM ROC
A6	᧶	KHMER SYMBOL PRAM-MUOY ROC
A7	᧷	KHMER SYMBOL PRAM-PII ROC
A8	᧸	KHMER SYMBOL PRAM-BEI ROC
A9	᧹	KHMER SYMBOL PRAM-BUON ROC
AA	᧺	KHMER SYMBOL DAP ROC
AB	᧻	KHMER SYMBOL DAP-MUOY ROC
AC	᧼	KHMER SYMBOL DAP-PII ROC
AD	᧽	KHMER SYMBOL DAP-BEI ROC
AE	᧾	KHMER SYMBOL DAP-BUON ROC
AF	᧿	KHMER SYMBOL DAP-PRAM ROC

Digit symbols for divination lore

B0	៰	KHMER SYMBOL LEK ATTAK SON
B1	៱	KHMER SYMBOL LEK ATTAK MUOY
B2	៲	KHMER SYMBOL LEK ATTAK PII
B3	៳	KHMER SYMBOL LEK ATTAK BEI
B4	៴	KHMER SYMBOL LEK ATTAK BUON
B5	៵	KHMER SYMBOL LEK ATTAK PRAM
B6	៶	KHMER SYMBOL LEK ATTAK PRAM-MUOY
B7	៷	KHMER SYMBOL LEK ATTAK PRAM-PII
B8	៸	KHMER SYMBOL LEK ATTAK PRAM-BEI
B9	៹	KHMER SYMBOL LEK ATTAK PRAM-BUON

Pali/Sanskrit extending sign

BA	ៜ	KHMER SIGN AVAKRAHA

Control character

BB	VI	KHMER VARIANT SIGN

Unassigned

BC - FF	<reserved>

Appendix Two

Comments on the contents of the existing Khmer character code table in ISO/IEC 10646-1:2000 and the Unicode Standard 3.1

1) The independent vowels (SRAK PENH TUA) ឣ (17A3) and ឤ (17A4) are included in the character table, but such characters do not actually exist in Khmer script. According to the Unicode Standard, they are used for transliteration of Pali/Sanskrit words. However, it is not an enough reason to include them, because they can be represented by the consonant អ (17A2), and by the consonant អ (17A2) + the vowel ា (17B6) respectively, if necessary.

2) ឨ (17A8) is included in the character table, but the Chuon Nath’s dictionary (“Dictionnaire Cambodgien”, 5^e édition, Institut Bouddhique, 1967-1968) specifically says that it is a ligature of ឧ (17A7) + ក (1780).

3) The independent vowel ឲ (17B2) is included in the character table, but it is a variant of ឱ (17B1).

4) Two inherent vowels (17B4) and (17B5) are included in the character table. In fact, such characters have never been used in Khmer and do not actually exist.

5) The dependent vowels (SRAK NISSAI) ◌ុំ and ◌ាំ are regarded not as single vowel signs but as combinations of NIKAHIT ◌ំ (17C6) and a vowel sign in ISO/IEC 10646-1 and the Unicode Standard. This is against the stance of the Chuon Nath’s dictionary.

6) Subscript consonants (COENG PYUNHCANA) are not assigned independent code points, but are instead represented by a control character ◌្ (17D2) plus the corresponding consonant from the character code table, based on the Indic (ISCII) *“VIRAMA MODEL”* (see “Khmer and Burmese Ad-Hoc Meeting Report”, ISO/IEC JTC 1/SC 2/WG 2 N1729, 1998-03-18).

   Behind this determination, there seems to be the idea that a subscript consonant is just a different glyph of its corresponding consonant. However there is more than that between them.

   First, a consonant can constitute an independent syllable by itself, but a subscript consonant cannot. In other words, if the former is a character in a narrow sense, the latter is a diacritic. Their relation is similar to that of an independent vowel (SRAK PENH TUA) and a dependent vowel (SRAK NISSAI) of the same pronunciation. As long as each of these vowels has its own code point, each of the two types of consonants should have an independent code point.

Second, it cannot be determined automatically in Khmer script whether a character code value for a consonant should be presented in a normal form or in a subscript form. This is not the case with Arabic script, where the presentation form of a character is automatically determined by its position and situation in a word. As long as the two have to be distinguished at the character code level, they are different character. Then different characters should have different code points according to one of the principles of ISO/IEC 10646-1 and the Unicode Standard.

7) The BATHAMASAT (17D3), (in Khmer: PATHAMASAT) is presumably included to represent the first August of leap year in lunar calendar, but we cannot find any code point assigned for the second August (TUTEYASAT).

8) An independent code point is assigned to ។ល។ (17D8), an abbreviation for the Khmer word meaning “et cetera”. Like “etc.” expressed by “e+t+c+.” in English, it can be written as a combination of “។+ល+។” (17D4+179B+17D4), so there is no need for a special code point to represent it. Furthermore, there are other ways of abbreviating this word, and it would be inconsistent to include only one of them in the character code table.

9) The same character or the same combination of characters often has more than one presentation form in Khmer (for example: ឧក and ឪ, ឱ and ឲ, ក៑ and ក̭, **ឥ** and **ឥ**, **ឦ** and **ឦ**, etc.). However no consistent way to deal with them can be found in the existing table.

10) As a consequence of some of the above decisions, the normal sequence of characters as used in the Chuon Nath’s dictionary has been violated, and this in turn presents unnecessary difficulties for sorting algorithms.