ISO
INTERNATIONAL ORGANIZATION FOR STANDARDIZATION
ORGANISATION INTERNATIONALE DE NORMALISATION

ISO-IEC JTC1/SC2/WG2
Multi-Octet Coded Character Set

ISO-IEC JTC1/SC2/WG2 _____
July 30, 1997

Title: Proposal for Encoding of the Khmer Script
Source: Maurice J Bauhahn
Status: Expert Contribution
Requested Action: Consideration by WG2 and UTC

This proposal draws heavily from the *Unicode Technical Report #1, Draft Proposal,* Copyright© 1992 Unicode, Inc., authored by Andy Daniels.

Khmer Script Encoding U+1400 to U+146F; The code positions specified in this proposal should not be used for any purpose other than evaluating this proposal.

Background

Cambodian, also known as Khmer, is the official language of Cambodia. Mutually intelligible dialects are also spoken in northeastern Thailand and the Mekong Delta region of Vietnam. While not itself an Indo-European language, much of the administrative, military and literary vocabulary of Khmer is borrowed from Sanskrit. With the advent of Theravada Buddhism at the beginning of the fifteenth century, Khmer began to borrow Pali words, and continues to use Pali as a major source of neologisms today. There is also much cross-borrowing between Thai and Khmer, as well as a relatively recent infusion of French words and a smattering of Chinese and Vietnamese loan-words in colloquial speech.

The Khmer script, called *a'saa kmae* (អក្សរខ្មែរ “Khmer letters”), as well as Thai, Lao, Burmese, Old Mon and others, are all descended from the Brahmi script of South India. The exact geological source, or possibly sources, has not been determined, but there is a great similarity between the earliest inscriptions in the region and the Pallawa script of the Coromandel coast of India. Various minority languages in Cambodia are being transcribed in the Khmer script, also.

Structurally, the Khmer script stays very close to its southern Brahmi origins. There is a set of 35 consonants, each with an inherent vowel sound.

In Khmer the inherent vowel sound is long, however certain words derived from Pali have inherent short vowel sounds. Consideration is given to having a code, U+1434 KHMER VOWEL INHERENT A, to represent this inherent short vowel sound. The inherent long vowel sound, U+1435 KHMER VOWEL INHERENT AA, is the default; hence it does not normally have to be in the text stream (but may be automatically and temporarily added for sorting purposes). Neither inherent form is normally visible. Additional characters are placed before, above, below or after the consonants (or a combination thereof) to indicate vowels other than the inherent one.

Consonant clusters are represented by conjunct consonants, where the first (base) consonant of the cluster maintains its full form and succeeding consonants are written as prescripts, subscripts, or postscripts. In this encoding the concept of a *wirama* ◌្ is borrowed from ISCII to convert the character following it into a conjoined character. That application has not been used in Khmer, but is standing in for the Khmer concept of ‘foot’ [subscript] (ជើង *chung*), which applies chiefly to consonants (all except one, ឡ), but also rarely to a handful of independent vowels:

បង្ឧផន   ផ្ងធ   វជ្យទរ្គិន៑   ហ្ឬទ័យ

There is also a representation of the rarely used *viriam* ◌៑ which indicates that a final consonant is part of the word preceding it...not stand-alone or part of the succeeding word. *Viriam* has in the past been represented by the symbol which herein represents *wirama* (not that *wirama* will be displayed...it only affects the display of the character following it). The following table paralleling the encoding table, illustrates the subjoined form taken by characters preceded by a *wirama*:

	140	141	142	143	144	145	146	147
0	◌្ក	◌្ថ	◌្ហ					
1	◌្ខ	◌្ទ						
2	◌្គ	◌្ធ	◌្អ					

	140	141	142	143	144	145	146	147
3	◌្ឃ	◌្ន						
4	◌្ង	◌្ប						
5	◌្ច	◌្ផ						
6	◌្ឆ	◌្ព						
7	◌្ជ	◌្ភ	◌្ឧ					
8	◌្ឈ	◌្ម						
9	◌្ញ ◌្ញ	◌្យ						
A	◌្ដ	◌្រ						
B	◌្ឋ	◌្ល	◌្ឫ					
C	◌្ឌ	◌្វ						
D	◌្ឍ	◌្ស						
E	◌្ណ	◌្ហ ◌្ហ						
F	◌្ត	◌្អ	◌្ឯ					

The Khmer language has a much more rich set of vowels than the Indo-Aryan languages for which the ancestral script was used. By the same token, there is a much smaller set of consonant sounds. The script is adapted to the Khmer language by adding extra vowel signs and various diacritic marks, and by using the choice of consonant as well as of vowel signs to determine the particular vowel sound represented. Thus most vowel signs do not have a single value but must be interpreted in the context of the associated consonant. This is similar to the situation in Thai and Lao, where different consonant symbols have the same sound but encode different tones.

There are two basic styles of script in modern Khmer, each with two major variations. They are the *a’saa trait* (“slanted script”) and the *a’saa muul* (អក្សរមូល “round script”). The “standing” variant of the slanted script is chosen here as representative.

Representation:

The Khmer script follows the model of Devanagari and other Indic scripts. The basic unit is the syllabic cluster in the following canonical order consisting of a series of consonants (or rarely independent vowels) separated by an invisible WIRIAM (◌្ 1450), followed by an optional vowel, followed by none or one of the quality marks MUSEKATOAN (◌៉ 1449) and TRUYSAP (◌៊ 144A) and/or by none or one of the diacritics. Note that second (or third) consonants in a cluster have variant forms and are usually placed under the first consonant with right margins aligned. In the case of U+141A KHMER LETTER RAW, the conjunct form is placed before the primary consonant. The conjunct forms of ឃឈឍយមស include a subscript component but also extend to the right of the primary consonant. A minimal syllabic cluster consists of one consonant or one independent vowel. A maximal syllabic cluster would consist of one consonant, two WIRIAM+consonant pairs, one vowel, one quality mark, and one diacritic. It is the responsibility of the display software to place conjoined forms in prefix, first or second subscript, or suffix position; to position component glyphs of vowels, to create ligatures (especially the obligatory បា or បោ or បៅ forms which combine ប+ា or ប+ េ◌ា or ប+ េ◌ៅ respectively), and to avoid overlap of characters. For example, the word *knyom*, ខ្ញុំ “I”, is coded as the string: U+1401; U+1452; U+1409; U+143B; U+1446. Display software needs to change the combination U+1452; U+1409; from ញ to ◌្ញ as well as shifting U+143B to second subscript position ◌ . Another example is the word កញ្ជ្រួច (*kaanychruach*)៖ This should be encoded U+1400; U+1409; U+1452; U+1407; U+1452; U+141A; U+143D; U+1405. Note the change of the form of the primary consonant U+1409 when a conjoined consonant is attached, the lengthening of the conjoined U+141A, and the downward shift of the vowel U+1405.

In cases where there is already some other superscript in the cluster, either of the two quality marks is written as the subscript symbol *kbiah kraom*, which looks much like U+143B VOWEL SIGN O. This vowel sign is not to be used for that purpose. It is the responsibility of the presentation software to select the correct appearance of the shifter. For example, *sii*, ស៊ី ⇒ សុី "to eat," should be coded as 141F+1436+1448, not as 141F+1438+144A.

KHMER SIGN RAWBAT (◌៌ 144C) historically corresponds to the Devanagari *repha*, that is, to an initial /r/ រ . It has lost this function in Khmer and instead is considered a simple diacritic similar to KHMER SIGN TOANDAKHIAT ◌៍ in both reading and sorting. There are also many cases of consonant clusters with initial /r/ that should be written with a full KHMER LETTER RAW រ and not a KHMER SIGN RAWBAT ◌៌ , so a separate character is provided for it.

Khmer writing does not normally separate words with white space as European languages do. If it is desirable to represent word boundaries in the text stream, for example, for use by automatic line layout algorithms, U+200B, ZERO WIDTH SPACE, should be used. Phrases in Khmer are separated by white space: U+0020, SPACE. A second type of aesthetic white space could be used to fully justify a line of Khmer text while only mildly reflecting on the grammatical phrasing of the text. A separate proposal will offer this SOFT SPACE entity since it could have wider application than in this one script.

Two relatively rare symbols that originated in the Khmer script are included here. They are *pnek moan*, U+1457 KHMER PHNEK MUAN ៙ "cock's eye" and *komout* , U+1459 KHMER KOMOUT ៚ "cow pee". They are identical in form and function to THAI CHARACTER FONGMAN (0E4F) and THAI CHARACTER KHOMUT (0E5B), respectively. It was thought inappropriate to use the characters with a Thai name when these originated in Khmer, given the cultural sensitivity of the Khmer.

Block Structure:

U+1400 to U+1422	Consonants
U+1423 to U+1424	Unassigned/reserved
U+1425 to U+1433	Independent Vowels
U+1436 to U+1445	Dependent Vowels
U+1446 to U+1448	Diacritics
U+1449 to U+144A	Quality Marks

U+1448 to U+1450	Diacritics
U+1451	Viriam
U+1452	Wirama
U+1453	Unassigned/reserved
U+1454 to U+1459	Symbols and Punctuation
U+145A	Unassigned/reserved
U+145B	Symbols and Punctuation
U+145C	Currency
U+145D to U+145F	Unassigned/reserved
U+1460 to U+1469	Digits
U+146A to U+146F	Unassigned/reserved

Issues:

The independent vowel LETTER AO TYPE 2 ឲ is a variant form of LETTER AO TYPE 1 ឱ. LETTER AO TYPE 2 apparently occurs only in the combination *aoy*, ឲ្យ "give", or in the word *auknya* (ឱកញ៉ា page 1860, Chhuan Nath's *Dictionnaire Cambodgien*).

In 1996 a committee of leading Khmer linguists in Cambodia clarified the status of three signs that have often been confused with vowels: KHMER SIGN NIKAHAT ◌ំ KHMER SIGN REAHMUK ◌ះ and KHMER SIGN YUKALEAPINTU ◌ៈ These are not vowels, but are often singly combined with vowels to change their pronunciations.

Words containing any one of these three signs (or other diacritics) are sorted after words otherwise identical but lacking these three signs (or other diacritics). Sorting is first on primary consonant by cluster (or inherent consonant of independent vowels, usually អ, but រ for ឫឬ and ល for ឭឮ), second on first conjunct consonant (or inherent consonant of independent vowels) of cluster, third on second conjunct consonant (or inherent consonant of independent vowels) of cluster, fourth on vowel (or inherent vowel equivalent of independent vowels) of cluster, and lastly (note the shift from *cluster*) *words* which contain quality marks or diacritics follow *words* that do not (arbitrarily in encoded order).

The vowel encoding takes an ISCII-like approach, coding as single characters each vowel that consists of two or more disjoint glyphs.

Two positions, U+1423 and U+1424, have been left unassigned pending resolution of discussions on their utility to transliterate independent vowels belonging to Pali words within Khmer contexts.

The proposed encoding (92/10/03; rev 92/11/25; rev 97/07/29) is:

	140	141	142	143	144	145	146	147
0	ក	ថ	ហ	ឮ	េ◌ៀ	◌័	០	
1	ខ	ទ	ឡ	ឱ	េ◌	◌៑	១	
2	គ	ធ	អ	ឯ	ែ◌	◌្	២	
3	ឃ	ន		ឳ	ៃ◌		៣	
4	ង	ប		◌ា់	េ◌ា	។	៤	
5	ច	ផ	ឥ	◌ា	េ◌ៅ	៕	៥	
6	ឆ	ព	ឦ	ា	◌ំ	៖	៦	
7	ជ	ភ	ឧ	◌ិ	ះ	ៗ	៧	
8	ឈ	ម	ឨ	◌ី	ៈ	។ល ។	៨	
9	ញ	យ	ឩ	◌ឹ	◌៉	៙	៩	
A	ដ	រ	ឪ	◌ឺ	◌៌			
B	ឋ	ល	ឫ	◌ុ	◌់	៚		
C	ឌ	វ	ឬ	◌ូ	◌៏	៛		
D	ឍ	ឝ	ឭ	◌ួ	◌៍			

	140	141	142	143	144	145	146	147
E	ណ	ឞ	ឮ	◌ើ	◌៎			
F	ត	ស	ឯ	◌ឿ	◌៏			

Khmer Character Names

Consonants

1400 KHMER LETTER KAA
1401 KHMER LETTER KHAA
1402 KHMER LETTER KAW
1403 KHMER LETTER KHAW
1404 KHMER LETTER NGAW
1405 KHMER LETTER CAA
1406 KHMER LETTER CHAA
1407 KHMER LETTER CAW
1408 KHMER LETTER CHAW
1409 KHMER LETTER NYAW
140A KHMER LETTER DAA
140B KHMER LETTER RETROFLEX THAA
140C KHMER LETTER DAW
140D KHMER LETTER RETROFLEX THAW
140E KHMER LETTER NAA
140F KHMER LETTER TAA
1410 KHMER LETTER THAA
1411 KHMER LETTER TAW
1412 KHMER LETTER THAW
1413 KHMER LETTER NAW
1414 KHMER LETTER BAA
1415 KHMER LETTER PHAA
1416 KHMER LETTER PAW
1417 KHMER LETTER PHAW
1418 KHMER LETTER MAW
1419 KHMER LETTER YAW
141A KHMER LETTER RAW
141B KHMER LETTER LAW
141C KHMER LETTER WAW
141D KHMER LETTER SHAA
Sanskrit
141E KHMER LETTER SSAA
Sanskrit
141F KHMER LETTER SAA
1420 KHMER LETTER HAA

1421 KHMER LETTER LAA
1422 KHMER LETTER QAA
glottal stop

Independent Vowels

1425 KHMER LETTER E
1426 KHMER LETTER EY
1427 KHMER LETTER O
1428 KHMER LETTER OK
1429 KHMER LETTER OU
142A KHMER LETTER UW
142B KHMER LETTER RIK
142C KHMER LETTER RII
142D KHMER LETTER LIK
142E KHMER LETTER LII
142F KHMER LETTER AE
1430 KHMER LETTER AY
1431 KHMER LETTER AO TYPE 1
1432 KHMER LETTER AO TYPE 2
1433 KHMER LETTER AW

Dependent Vowels

1434 KHMER VOWEL INHERENT A
1435 KHMER VOWEL INHERENT AA
1436 KHMER VOWEL SIGN AA
1437 KHMER VOWEL SIGN E
1438 KHMER VOWEL SIGN EY
1439 KHMER VOWEL SIGN U
143A KHMER VOWEL SIGN UI
143B KHMER VOWEL SIGN O
x kbiah kraom
143C KHMER VOWEL SIGN OU
143D KHMER VOWEL SIGN UA
143E KHMER VOWEL SIGN AU
143F KHMER VOWEL SIGN IU
1440 KHMER VOWEL SIGN IE
1441 KHMER VOWEL SIGN EI
1442 KHMER VOWEL SIGN AE
1443 KHMER VOWEL SIGN AY
1444 KHMER VOWEL SIGN AO
1445 KHMER VOWEL SIGN AW

Diacritics

1446 KHMER SIGN NIKAHAT
= damla
1447 KHMER SIGN REAHMUK
= wihsacini
1448 KHMER SIGN YUKALEAPINTU
= coc pi

Quality Marks

1449 KHMER SIGN MUSEKATOAN
= tmin kandao

144A KHMER SIGN TRUYSAP
=sok kaw

Diacritics

144B KHMER SIGN BANTOC
= reahsannya
144C KHMER SIGN RAWBAT
= rephea
144D KHMER SIGN TOANDAKHIAT
= samlap
= patdesaet
144E KHMER SIGN KAKABAT
= caung kaek
144F KHMER SIGN AHSDA
= leik prambuy
1450 KHMER VOWEL SIGN SANYOK SANNYA
=phuat leu
1451 KHMER SIGN VIRIAM
Virama
1452 KHMER SIGN WIRAMA

Symbols and Punctuation

1454 KHMER KHAN
full stop
ellipsis
1455 KHMER BARIYAOSAN
end of section
1456 KHMER CAMNOC PI KUH
x (division sign -> 00F7)
x (tibetan comma -> 1038)
colon, semicolon
1457 KHMER LEIK TO
= amendit sannya
repetition sign
1458 KHMER BEYYAL
= leh
continuation sign
1459 KHMER PHNEK MUAN
= kakodney
list bullet
145B KHMER KOMOUT
document end

Currency

145C KHMER CURRENCY SYMBOL RIEL

Digits

1460 KHMER DIGIT ZERO
1461 KHMER DIGIT ONE
1462 KHMER DIGIT TWO
1463 KHMER DIGIT THREE
1464 KHMER DIGIT FOUR
1465 KHMER DIGIT FIVE

1466 KHMER DIGIT SIX
1467 KHMER DIGIT SEVEN
1468 KHMER DIGIT EIGHT
1469 KHMER DIGIT NINE

References:

Institut Bouddhique. *Dictionnaire Cambodien*. Phnom Penh, 1967. (also known as Chhun Nath's Dictionary)

"Decisions of the secretariat of the National Higher Education Task Force regarding encoding of the Khmer language into Unicode/computer".
14 August 1996.