ISO
INTERNATIONAL ORGANIZATION FOR STANDARDIZATION
ORGANISATION INTERNATIONALE DE NORMALISATION

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

ISO-IEC JTC1/SC2/WG2 _____
March 17, 1998

Title:	Proposal for Encoding 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+146E; The 103 code positions actually specified in this proposal should not be used for any purpose other than evaluating this proposal. The naming transliterations used in this proposal are phonetic...based loosely on the American Library Association-Library of Congress romanization tables (1997). Variations from these relate to the restricted symbol set, pronunciation changes in context, or established norms. Unfortunately no scheme is ideal.

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 *qaksakhmaer* (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 AQ, to represent this inherent short vowel sound for scholarly purposes (but not conventional text). 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 phonetic or sorting purposes). Neither inherent form is normally visible.

Alternative glyphs are placed before, above, below, or after (or a paired combination thereof) the consonants to indicate a single vowel character 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 *virama* ្ is borrowed from ISCII to denote that the character preceding it drops its vowel and accepts a conjoined character. A glyph for that purpose is not used in Khmer, but the strong verbal Khmer concept of 'foot' [subscript] (ជើង JOENG) is *always* connected in Khmer spelling with conjoined entities. The proposed code for KHMER SIGN JOENG is U+1452. Conjoined characters are chiefly consonants (all except one, ឡ), but also rarely a handful of independent vowels (the number of which is very difficult to guarrantee):

បង្ខផន ផ្ងធ វជ្យទរិន្ក¯ ហ្ឫទ័យ

There is also a representation of the rarely used KHMER SIGN VIRIAM U+1451 ៑ 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 KHMER SIGN JOENG (not that JOENG will be displayed...it only affects the display of the character following it). The following table paralleling the encoding table, illustrates the conjoined forms (of that subset currently documented; U+141E is mentioned in the ALA-Library of Congress romanization standard [although the author of this proposal has not yet seen it used in a Khmer word]) taken by characters preceded by a *joeng*:

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

	140	141	142	143	144	145	146	147
C	◌្ឈ	◌្ហ						
D	◌្ញ	◌្ក						
E	◌្ណ	◌្ម ◌្ផ						
F	◌្ដ	◌្ប	◌្ង					

The Khmer language has a richer set of vowels than the Indo-Aryan languages for which the ancestral script was used. By the same token, there is a 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 *qaksa jrieng* (អក្សរជ្រៀង “slanted script”; អក្សរឈរ “standing script”) and the *qaksa muul* (អក្សរមូល “round script”; ខម). 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 one or more consonants (or rarely independent vowels) separated (if they number two or three) by an invisible KHMER SIGN JOENG (◌្ U+1452), followed by an inherent or explicit vowel, followed by none or one of the quality marks KHMER SIGN MUUSIKATOAN (◌៉ U+1449) and KHMER SIGN TRUYSAP (◌៊ U+144A) and/or by none or one of the diacritics (U+1446, U+1447, U+144B through U+1450). Note that second (or third) characters in a syllabic cluster have variant forms and are usually placed under the first consonant with right margins aligned. All but one consonant (U+1421) in the range U+1400 to U+1422 can have a conjunct form in the first subscript position. The following characters are known to also have forms in the second subscript position: conjuncts of មរសហ and part of the conjunct of ញ under itself and the vowels ◌ុ ◌ូ ◌ួ ◌ៀ ◌ឿ. These vowels may even occur in third subscript position, depending on the shape of the conjunct of រ. In the case of U+141A KHMER LETTER RO, the conjunct form is placed before the primary consonant. The conjunct forms of ឃឈឍយឞស 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 JOENG+conjunct pairs, one dependent vowel (possibly composed of two glyphs or even joined to the consonant in a ligature), one quality mark, and one diacritic (although possibly U+1446 to U+1448 could coexist with other diacritics). 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 high បា or បោ or បៅ forms which combine ប+ា or ប+ ◌ោ or ប+ ◌ៅ respectively), and to avoid overlap of characters. High ligature forms join ◌ា or ◌ោ or ◌ៅ to កគឃចឆឈឍតធភមយរលសឡ. Low ligature forms join ◌ា or ◌ោ or ◌ៅ to ជដ. 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 (*kaanyjhruac*)៖ 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 from ញ to ញ when a subjoined consonant is attached, the lengthening of the conjoined U+141A (some attractive glyph designs bring the bottom sweep of this conjoined form all the way to line up with the right side of the supporting consonant, vertically positioned between the first subscript and the vowel...effecting for the vowel a third subscript layer), 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 U. 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 ⇒ ស៊ី "to eat," should be coded as 141F+144A+1448, not as 141F+1438+143B.

KHMER SIGN RAWBAT (◌៌ U+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 TOANDAGHIAT ◌៍ in both reading and sorting. There are many consonant clusters with initial /r/ that should be written with a full KHMER LETTER RO U+141A រ, not a KHMER SIGN ROWBAT U+144C ◌៌ , 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 should be used to fully justify a line of Khmer text while mildly reflecting on the grammatical phrasing of the text. A careful reading of the sense of the phrase is needed to determine appropriate placement of this largely aesthetic white space, so algorithmic generation is not appropriate. 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 U+1459 KHMER PHNAEK MUAN ៙ “cock's eye” and U+145B KHMER GOOMUUT ៚ “cow pee”. They are identical in form and function to THAI CHARACTER FONGMAN (U+0E4F) and THAI CHARACTER KHOOMUUT (U+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	Transliterations of Independent Vowels
U+1425 to U+1433	Independent Vowels
U+1434 to U+1425	Inherent 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	Joeng
U+1453 to U+145A	Symbols and Punctuation
U+145B	Currency
U+145C	Punctuation
U+145D to U+145F	Unassigned/reserved

U+1460 to U+1469 Digits
U+146A to U+146F Unassigned/reserved

Issues:

The KHMER INDEPENDENT VOWEL QO TYPE 2 U+1432 ឲ is a variant form of KHMER INDEPENDENT VOWEL QO TYPE 1 U+1431 ឱ. KHMER INDEPENDENT VOWEL QO TYPE 2 apparently occurs only in the combination *oy*, ឲ្យ "give", or in the word *oknya* (ឲកញ៉ា page 1860, Juan Nat'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 NIGAHIT ◌ំ KHMER SIGN REAHMUK ◌ះ and KHMER SIGN YUUKALEAPINTU ◌ៈ These are not vowels, but are often singly combined with vowels to change their pronunciations. They are combined with at least eleven dependent vowels.

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 after their consonant each vowel that consists of one or two disjoint glyphs (part of which may precede the consonant in display position).

Two positions, U+1423 and U+1424, have been assigned characters exclusively for transliteration of analogous Indic independent vowels in linguistic, religious, or scholarly publications to preserve reversability of transliteration between languages. They should not be used in place of the analogous consonant/consonant-vowel-ligature in contemporary literature.

One position, KHMER INDEPENDENT VOWEL QYK U+1428, contains an independent vowel of historic interest that is not currently in use.

Thel Thong has researched some of the historic ways of writing dates in Khmer. Two examples of such dates are:

ថ្ងៃ ៦ ᧲ ៩ ឆ្នាំ (day 6 [Friday], 2nd day of the waned moon of the ninth lunar month of year....; a non-leap year date.

ថ្ងៃ ១ ᧢ (᧠) ឆ្នាំ (day 1 [Sunday], 2nd day of the waxed moon of the eighth lunar month [Pathameasath, បឋមាសាឍ] of year....; a leap year date.

ថ្ងៃ ៧ ᧵ (៨៨) ឆ្នាំ (day 7 [Saturday], 5th day of the waned moon of the eighth lunar month [Tutiyasath, ទុតិយាសាឍ] of year....; a leap year date.

Some minority languages of Cambodia are being written in the Khmer script. However additional codes will be needed to represent consonant sounds not present in Khmer. For example, Khmer lacks unimploded first and second register ‘b’, unimploded first and second register ‘d’, first and second register ‘g’ and ‘j’, first and second register imploded ‘y’, and imploded ‘g’. These are all present in Krung. Although coding points are not proposed here, I recommend that the Unicode consortium reserve at least 16 more points for the eventual standardization of these and other related characters.

Additional characters used frequently in Khmer text can be found elsewhere in Unicode/ISO10646:

U+0020	SPACE (phrase break)
U+2027	WORD-BREAK SPACE (invisible word break)
U+????	DISCRETIONARY PHRASE SPACE
U+2007	FIGURE SPACE (thousands separator)
U+200D	ZERO-WIDTH JOINER (to join numbers together and with U+1453 or U+1454 for subscripting/superscripting)
U+200C	ZERO-WIDTH NON-JOINER (prevent ligature formation)
U+00AB	LEFT POINTING GUILLEMET (quotation) «
U+00BB	RIGHT POINTING GUILLEMET (quotation) »
U+003F	QUESTION MARK ?
U+201D	QUOTATION MARK, DOUBLE COMMA (quotation) ”
U+201C	QUOTATION MARK, DOUBLE TURNED COMMA (quotation) “
U+2032	PRIME (mathematic expressions) ′
U+301E	PRIME, DOUBLE (mathematic expressions) ″
U+002B	PLUS SIGN +
U+2212	MINUS −
U+00D7	MULTIPLICATION SIGN ×
U+00F7	DIVISION SIGN ÷
U+003D	EQUALS SIGN =
U+2010	HYPHEN -
U+00AD	SOFT HYPHEN -
U+0021	EXCLAMATION MARK !
U+002C	COMMA , (unfortunately used by different Khmer users as either thousands separator or decimal mark)

U+0030 TO U+0039	DIGIT ZERO through DIGIT NINE 0123456789
U+002E	PERIOD . (unfortunately used by different Khmer users as either thousands separator or decimal mark)
U+0025	PERCENT SIGN %
U+0078	BRACKET, OPENING CURLY {
U+007D	BRACKET, CLOSING CURLY }
U+0028	PARENTHESIS, OPENING (
U+0029	PARENTHESIS, CLOSING)
U+005B	SQUARE BRACKET, OPENING [
U+005D	SQUARE BRACKET, CLOSING]
U+25CC	DOTTED CIRCLE ◌ (display conjuncts, independent vowels and diacritics in isolation)

References:

Institut Bouddhique. *Dictionnaire Cambodien*. Phnom Penh, 1967. (also known as Chuon 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.

Thel Thong. សិលាចារឹកព្រះគោ Khmer Students Association of Victoria. Quarterly Newsletter. Vol. 3, No. 2, June, 1977, pp. 12-15.

American Library Association - Library of Congress romanization tables: transliteration schemes for non-Roman scripts/approved by the Library of Congress and the American Library Association; tables compiled and edited by Randall K. Barry. - 1997 ed. ISBN 0-8444-0940-5