**Universal Multiple-Octet Coded Character Set**
**(UCS)**

ISO / IEC JTC 1 / SC 2 / WG 2 **N2459**
DATE: 2002-05-14

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

DOC TYPE: National Body Contribution
TITLE: Cambodian stance regarding Khmer code table in Unicode 3.2
SOURCE: Cambodian Committee for Standardization of Khmer Characters in Computers, National Information Communications Technology Development Authority *
STATUS: For discussion at 42nd WG2 meeting in Dublin
DISTRIBUTION: ISO / IEC JTC 1 / SC 2 / WG2 and UTC
MEDIUM: Electronic
NO. OF PAGES: 6

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

Almost exactly one year ago in letters dated 30 and 31 May 2001 (SC2 N3571) Cambodia registered its official objection to the Khmer code table adopted by Unicode, requesting rescission and replacement. Our position was subsequently spelled out in more detail in N2380R, N2406 and N2458.

Cambodian delegations were able to attend the 41st WG2 in Singapore and UTC90 in San Jose to speak to our case and to hear responses to it. Following such deliberations, and consultation with expert linguists, we remain if anything even more convinced that the Khmer code table adopted by Unicode is not only inefficient but does not accord with Cambodian perceptions of their language. Furthermore, we maintain our position that neither Unicode nor ISO / IEC followed correct procedures in discussing, adopting and publishing this standard without involving or even informing the registered Cambodian ISO subscriber member.

At the same time, however, we have been given firm indications that our request for rescission of the Khmer code table and its replacement, or even the deprecation of the artificial COENG character (U+17D2) and addition of explicit code points for the subscript characters will not be accepted. Evidently Unicode and ISO / IEC place the principle of not changing the published standard higher than the principles of faithfulness to the script and due process.

Faced with this situation, the Cambodian National Body has the choice of going ahead with its own national standard that does not follow Unicode, and requesting an additional block of characters in Unicode for individual subscript characters (whether

canonically equivalent or not), or accepting to work with the current Khmer code table in the face of *force majeure*. In its meeting on 7 May 2002 the Committee for Standardization of Khmer Characters in Computers decided with regret to adopt the Khmer code table in Unicode 3.2. This decision was taken principally because the Committee felt that Cambodia could not pay the price of further delay in implementation of its script nor the redundancy and incompatibility that would result from adoption of an incompatible national standard.

We appreciate the action taken in Unicode 3.2, in response to the Cambodian request at UTC90, to spell out explicitly the character names and code points required for individual subscript characters, adding some explanation of the artificial approach followed by Unicode, and discouraging five characters prior to possible deprecation and addition of a full set of characters to represent lunar dates.

Furthermore, we note with considerable appreciation the letter from Unicode President Mark Davis dated 25 April 2002 (see below) regretting the course followed by Unicode in defining the Khmer code table.

It is our request that the essence of his letter be incorporated into future versions of Unicode to set the record straight. We consider this necessary in particular for future generations of Cambodians who would otherwise undoubtedly criticise the Cambodian National Body for allowing this situation of such an artificial and inefficient coding to prevail. We would suggest inclusion of the following wording:

"In the words of Unicode President Mark Davis in a letter dated 25 April 2002

> As a result of our discussions we have learned that:
>
> 1. The encoding approach taken for the representation of the Khmer subscript letters in Unicode, the so-called 'virama model', is not the preferred approach of the Cambodian National Body or of Khmer linguistic experts, and is at odds with the way the Khmer script is perceived and taught in Cambodia.
> 2. A number of characters were added to the encoding of the Khmer script, which upon receipt of further input from the Cambodian National Body, now appear to have been clear mistakes. Those characters cannot properly be considered to be a part of Khmer script.
> 3. A number of symbols and other characters used in the representation of the Khmer script were overlooked in the encoding.
>
> The Unicode Consortium acknowledges and regrets that over the last several years, and especially during crucial periods when the decisions about development of the Khmer script encoding proposals were being made, that insufficient efforts were made to maintain full communication and consultation with all interested parties in Cambodia. This has resulted in the current unfortunate situation where all interested parties now have to deal with a less-than-optimal outcome with regard to Khmer encoding."

Furthermore, to avoid the possibility of a recurrence of this situation, we propose that WG2 and UTC adopt the following resolution:

“That every effort will be made to include the participation of appropriate official national bodies in future proposals to encode additional characters and scripts in Unicode. Any proposals are to be forwarded with a request for comment to appropriate official national bodies when they are first tabled for discussion, and the final position of the appropriate official national bodies shall be circulated to all members prior to balloting.”

Only by making a genuine effort to include active participation by those who read and write the script under consideration can Unicode in the future hope to avoid being seen as culturally and technologically insensitive to smaller and poorer countries which have not yet had their scripts encoded and who find it hard to marshal the financial and technical resources to be spontaneous and constant participants in the standard-setting process, thereby leaving the initiative in the hands of outsiders, however well-meaning they may be.

This past year has been a valuable learning experience for the Committee for Standardization of Khmer Characters in Computers and we would like to express our appreciation for those who helped us, and for those on the other side of the debate who treated our position with respect.

The Unicode Consortium
P.O. Box 391476
Mountain View, CA 94039-1476
USA

Tel: (650) 693-3921
Fax: (650) 693-3010
Email: info@unicode.org
http://www.unicode.org

Dr. Sorasak Pan
Under Secretary of State
Russian Federation Blvd., Room C12
Phnom Penh, Cambodia

April 25, 2002

Dear Dr. Pan,

The Unicode Consortium wishes to thank the Cambodian National Body for sending a delegation to the last meeting of the Unicode Technical Committee in Mountain View, California to discuss problems and issues apparent in the handling of the encoding of the Khmer script in the Unicode Standard and in the synchronized International Standard ISO/IEC 10646. The meeting and discussions were frank, cordial, and extremely informative.

As a result of our discussions we have learned that:

1. The encoding approach taken for the representation of the Khmer subscript letters in Unicode, the so-called "virama model", is not the preferred approach of the Cambodian National Body or of Khmer linguistic experts, and is at odds with the way the Khmer script is perceived and taught in Cambodia.

2. A number of characters were added to the encoding of the Khmer script, which upon receipt of further input from the Cambodian National Body, now appear to have been clear mistakes. Those characters cannot properly be considered to be a part of the Khmer script.

3. A number of symbols and other characters used in the representation of the Khmer script were overlooked in the encoding.

The Unicode Consortium acknowledges and regrets that over the last several years, and especially during crucial periods when the decisions about development of the Khmer script encoding proposals were being made, that insufficient efforts were made to maintain full communication and consultation with all interested parties in Cambodia. This has resulted in the current unfortunate situation where all interested parties now have to deal with a less-than-optimal outcome with respect to the Khmer encoding.

The Unicode Consortium sincerely regrets the impact that this may have on the perception of the Khmer script, as well as any confusion that may result for those who are implementing the Khmer script on the basis of the Unicode Standard. The Unicode Consortium has taken a number of steps to provide for the best rectification feasible for the mistakes that have been made in the Khmer script encoding.

These steps include:

1. Extension of a formal liaison relationship offer to the Cambodian National Body to ensure the closest possible interaction with IT experts in Cambodia (see attached information)
2. Approval of the addition of forty-five characters (digit symbols for divination symbols, lunar month symbols, the KHMER SIGN ATTHACAN, and voicing signs for the Krung dialect of Brao)
3. Initiation of the deprecation process for six questionable characters.
4. Clarification of Khmer script in the Unicode Standard Version 3.2, which was released at the end of March 2002. This includes:
    a. text explaining the proper linguistic analysis of Khmer, including the unitary status of subjoined consonants
    b. warnings about the status of the six characters that are candidates for deprecation
    c. code charts showing the encoding of subjoined consonants

(For more information, see <http://www.unicode.org/unicode/reports/tr28/#9_15_khmer>)

The Unicode Consortium wishes to express deep appreciation to the members of the delegation from Cambodia. As a result of their visit, the Unicode Technical Committee is better informed about particular features of the Khmer script and how the Cambodian people perceive it. We hope that the mutual respect and collegiality that was so apparent at the recent UTC meeting will continue into the future. The Consortium looks forward to continuing to work with the Cambodian National Body on improving the Unicode Standard to meet the particular requirements of Cambodian users so that Khmer script may be supported in Unicode-based software.

Best Regards,

Mark Davis
President,
The Unicode Consortium

Cc: CHEA, Sok Huor

***Liaison Membership***

Liaison Membership in the Unicode Consortium is offered (at no charge) to organizations that have interests in common with the Consortium, particularly in technical areas. The Unicode Consortium and the organization with which it has a liaison relationship each nominate a liaison representative.

Liaison representatives have primary responsibility for the interchange of information between the two parties. In particular, liaison representatives are charged with alerting the other member of the liaison relationship to events materially affecting that organization's area of interest or expertise.

Liaison members of the Unicode Consortium include both international groups (for example, The Internet Engineering Task Force (IETF) and the World Wide Web Consortium's I18N Working Group) and national bodies (for example, the High Council of Informatics of Iran (HCI) and the Technical Committee on Information Technology (TCVN/TC1) of Vietnam).

For more information on the current liaison memberships, see
<http://www.unicode.org/consortium/memblogo.html#5>