ISO/IEC JTC1/SC2/WG2 N4xxx
L2/12-350
2012-10-29

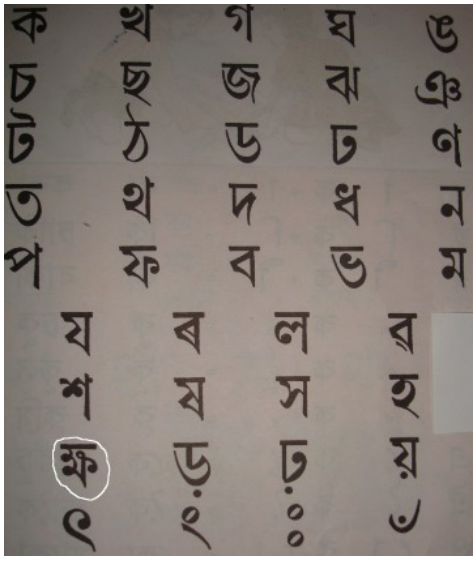
Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation internationale de normalisation
Международная организация по стандартизации

Doc Type: **Working Group Document**
Title: **Proposal for minor, non-character, additions to UCS addressing concerns from Assamese**
Source: **Martin Hosken**
Action: **For consideration by UTC & JTC1/SC2/WG2**
Date: **2012-10-29**

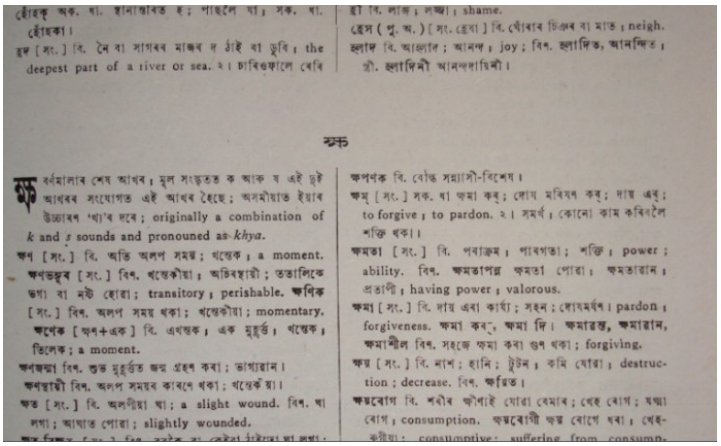
Introduction: The Bengali script is used for a number of languages, in particular Assamese. This proposal addresses a number of concerns that have come from the Assamese regarding the Unicode encoding of Bengali and Assamese. In particular this document incorporates the concerns as expressed in private communication from Jyotiprakash Tamuli.

Block Name: There seems to be some sensitivity towards the block name not mentioning Assamese. Given that the block name, once set, cannot change, it is proposed that a block header paragraph may be used to address the issue of the close relation between the Bengali and Assamese scripts. No text is proposed here for such a paragraph and those interested in crafting such a paragraph should propose some wording. It should be noted, that since such text is editorial, and not actually part of the standard, the editors do have the freedom to change such text should they see the need. In addition, such text should be considered appropriate by all readers of the standard, including speakers of Bengali and users in India who are not Assamese.

Ksha: Assamese has one conjunct that is specifically referenced in their alphabet. They want to be able to type this using one keystroke, and to give it its own sort position. There is no rendering difference between this individual character and the sequence `U+0995 BENGALI LETTER KA U+09CD BENGALI SIGN VIRAMA U+09B7 BENGALI LETTER SSA`.



2. Ex: Consonant chart showing ksha in the alphabet sequence


ক্ষ বর্ণমালার শেষ আখর ; মূল সংস্কৃতত ক আৰু ষ এই দুই আখৰৰ সংযোগত এই আখৰ হৈছে ; অসমীয়াত ইয়াৰ উচ্চাৰণ 'খ্যা'ৰ দৰে ; originally a combination of *k* and *ṣ* sounds and pronounced as *khya*.
ক্ষণ [সং.] বি. অতি অলপ সময় ; খন্তেক ; a moment.
ক্ষণভঙ্গুৰ [সং.] বিণ. খন্তেকীয়া ; অচিৰস্থায়ী ; তাত্ক্ষণিকে ভগা বা নষ্ট হোৱা ; transitory ; perishable.
ক্ষণিক [সং.] বিণ. অলপ সময় থকা ; খন্তেকীয়া ; momentary.
ক্ষণেক [ক্ষণ+এক] বি. এখন্তেক, এক মুহূর্ত, খন্তেক, তিলেক ; a moment.
ক্ষণজন্মা বিণ. শুভ মুহূর্তত জন্ম গ্ৰহণ কৰা ; ভাগ্যবান।
ক্ষণস্থায়ী বিণ. অলপ সময়ৰ কাৰণে থকা ; খন্তেকৰ।
ক্ষত [সং.] বি. অলপীয়া ঘা ; a slight wound. বিণ. ঘা লগা ; আঘাত পোৱা ; slightly wounded.
ক্ষপণক বি. বৌদ্ধ সন্ন্যাসী-বিশেষ।
ক্ষম্ [সং.] সক. বা ক্ষমা কৰ্ ; দোষ মৰিষণ কৰ্ ; দায় এৰ্ ; to forgive ; to pardon. ২। সমৰ্থ, কোনো কাম কৰিবলৈ শক্তি থকা।
ক্ষমতা [সং.] বি. পৰাক্ৰম ; পাৰগতা ; শক্তি ; power ; ability. বিণ. ক্ষমতাপন্ন ক্ষমতা পোৱা ; ক্ষমতাৱান, প্ৰতাপী ; having power ; valorous.
ক্ষমা [সং.] বি. দায় এৰা কাৰ্য্য ; সহন ; দোষমৰ্ষণ ; pardon ; forgiveness. ক্ষমা কৰ্, ক্ষমা দি। ক্ষমাৰ্হ, ক্ষমাৱান, ক্ষমাশীল বিণ. সহজে ক্ষমা কৰা গুণ থকা ; forgiving.
ক্ষয় [সং.] বি. নাশ ; হানি ; টুটন ; কমি যোৱা ; destruction ; decrease. বিণ. ক্ষয়িত।
ক্ষয়ৰোগ বি. শৰীৰ ক্ষীণাই যোৱা বেমাৰ ; ধেহ ৰোগ ; যক্ষ্মা ৰোগ ; consumption. ক্ষয়ৰোগী ক্ষয় ৰোগে ধৰা, ধেহ-ক্ষীয়া ; consumptive ; suffering from consump-
1. Ex: Dictionary showing head letter ksha

The proposed solution to this issue is not to encode a new character but to add a formal named sequence to cover the letter ksha:

```
BENGALI LETTER ASSAMESE KSSA;0995 09CD 09B7
```

Adding this named sequence will help implementers to be aware that in Assamese, such a sequence needs special handling. The issue of where such a sequence sorts should be addressed using sort tailoring. This is a mechanism that allows for language specific adjustments to the default sorting sequence. As to allowing such a sequence to be typed as a single keystroke, that is an implementation issue and should be raised with keyboard implementers.

Ra with diagonals: The two characters U+09F0 BENGALI LETTER RA WITH MIDDLE DIAGONAL and U+09F1 BENGALI LETTER RA WITH LOWER DIAGONAL are considered particularly inappropriately named. The characters are Assamese characters and so should be named based on their Assamese use. Since a character name cannot change, once set, there is no way to change the names. Instead we propose the addition of two character aliases in the names list:

09F0 = Assamese Ra

09F1 = Assamese Wa

Ru and Ruu: An expressed problem with rendering for Assamese regards the sequences U+09F0 BENGALI LETTER RA WITH MIDDLE DIAGONAL U+09C1 BENGALI VOWEL SIGN U, and U+09F0 BENGALI LETTER RA WITH MIDDLE DIAGONAL U+09C2 BENGALI VOWEL SIGN UU.

ASSAMESE			
		Currently rendered as	Correct conjunct form
1.	ৰ+ু	ৰু	ৰু
2.	ৰ+ূ	ৰূ	ৰূ

BENGALI		
		Correct conjunct form
1.	র+ু	রু
2.	র+ূ	রূ

Since the two base characters are used only for Assamese, implementers need to understand that for their solutions to adequately support Assamese, they must render the sequences as presented above. This is not an encoding issue, but one of educating font developers. It is assumed that newer versions of common fonts will address this issue if they have not already done so.

Conclusion: While there are concerns from the Assamese community regarding how Assamese is represented in Unicode, there are no technical problems that require any changes to the encoding itself. The above proposals are merely using mechanisms the UCS has for allowing communities to understand that the UCS does meet their technical needs in appropriate ways.