L2/11-308

A clear policy on encoding OM characters

Shriramana Sharma, jamadagni-at-gmail-dot-com, 2011-Jul-28

Recently I had submitted a proposal for Oriya OM L2/11-258 based on clear evidence from Oriya manuscripts etc that a ligated form of O + CANDRABINDU is used consistently apart from the mere sequence of O + CANDRABINDU rendered as a candrabindu placed on O.

Immediately afterwards, a proposal for a Bengali OM L2/11-275 has been submitted which however presents no evidence for an OM distinct from the sequence of O + CANDRABINDU. The South Asian Committee in L2/11-298 §4b has treated this proposal apparently on par with the Oriya proposal and recommended encoding this character without making any remarks on the technical justification of doing so.

Absence of technical justification for a distinct Bengali OM

I have noted on the Unicore list that there is no technical justification for encoding such a Bengali OM character which bears no visual distinction from the sequence of Bengali O + CANDRABINDU. The proposal itself gives no other justification. The fact that the syllable OM has unique religious significance (which is true even for this author) is not a justification for atomic encoding which should be decided only on technical bases.

Ubiquitous nature of Devanagari OM

The only semblance of justification in the proposal seems to be the remark that sometimes the OM in Bengali texts is presented similar/identical to Devanagari OM, possibly implying that since there is evidence of such glyphic variation it may be encoded separately. I must remark at this juncture that while researching the Oriya OM I was informed by native Oriya people that they also write the OM similar to what is currently encoded as the Devanagari OM. It is noted that the character encoded as the Gujarati OM is also glyphically highly similar/identical to the Devanagari OM. Evidently this written form for OM is quite prevalent all over North India.

However, given that this written form is already encoded at 0950, that is the codepoint/character to be used for representing that written form. Given that there are no mandatory word boundaries at script boundaries (except in Japanese/Latin text), there will be no text processing problems in using it in the middle of Bengali (or Oriya) text. In fact, given this, no separate Gujarati OM need have been encoded at all.

Problems with encoding Bengali OM without technical justification

The greatest problem with encoding such a Bengali (or any other script) OM which is glyphically identical to a sequence is that it would be confusable with the sequence giving rise to security risks in IDNs. Granted that there are other security measures to curb such risks, but there is no meaning to first encoding the character without any compelling technical reasons and then implementing security measures.

Accepting such a character would only give rise to other technically unjustified proposals for encoding similar sequences of O + CANDRABINDU or O + ANUSVARA or O + VOWELLESS-MA as atomic OM characters in the absence of any visual distinction.

In fact, already loose comments asking for the encoding of Telugu and Kannada OMs were made on the Unicore list without providing any evidence of distinct visual forms.

A clear policy on encoding OM characters

It is then clear that a clear technical policy on encoding OM characters is needed. **It is recommended that the policy be to encode an atomic OM character only when it is not visually identical to the default presentation of any sequence such as O + CANDRABINDU or O + ANUSVARA or O + VOWELLESS-MA (or such).** In fact, this is no new policy at all, for sequences are simply not encoded as atomic characters in Unicode!

If at all any Bengali, Telugu, Kannada or other script written forms for OM distinct from the pure sequences are found to be consistently used and proper attestation and evidence is provided for the same, such characters may be encoded and not otherwise.

A note in passing on Tibetan OM

It is noted that 0F00 TIBETAN OM is glyphically identical to the sequence 0F68 0F7C 0F7E. However, this need not affect the present policy as Tibetan is not currently classified along with Indic scripts (even though it is related to them). Feedback from Tibetan scholars may be sought as to the policy according to which this character should be used, whether its use should be discouraged like 0F73, 0F75 etc, whether it should be allowed in IDNs and so on.

The more serious matter of the Sharada OM

Returning to recognized Indic scripts, it is noted that the proposed Sharada encoding contains a Sharada OM at 111C4 which is identical to the sequence of O + CANDRABINDU in Sharada. I had first inquired on the Unicore list as to the justification of the acceptance of this atomic character but got no reply.

Later personal conversation with Anshuman Pandey, the author of the Sharada proposal, revealed more. It seems that a distinct ligated OM may also be attested in Sharada manuscripts, but this needs to be verified (as of date). I have asked Anshuman to look into the matter and verify it.

I however note that a ligated form of Sharada OM is already seen in the samples provided in my Oriya OM proposal L2/11-258 bottom of p 4. However, these samples are from recent OM collections and I would like this to be verified from manuscript sources which Anshuman has promised to look into.

Anshuman suggested (in personal conversation) that given that the Sharada OM exhibits two variants w.r.t. its candrabindu (see L2/09-074R2 pp 26,27 of PDF) — to wit that the candrabindu in OM may be either inverted [w.r.t. the commonly seen Indic form], or it may be upright — this may be taken as sufficient justification for atomic encoding.

However I feel this is not the case. Anshuman has already noted separately in his proposal (same page as above, immediately above the OM subheading) that the Sharada candrabindu *in general* is presented in both ways: while the inverted candrabindu is the more prevalent in Sharada, the upright form is also seen. In fact Anshuman's samples showing this show an upright candrabindu above a character which is certainly not an O. Therefore the variation of the candrabindu seen in the Sharada OM is not peculiar to the OM but is a general feature of the Sharada script.

Therefore pending verification of the evidence of a ligated form of Sharada OM, the character as it stands, being glyphically identical to the sequence of O + CANDRABINDU, cannot be justifiably encoded as an atomic character.

However, the Sharada proposal is at a very mature stage in the ISO process and it is probably not possible to suspend this character at this stage. At least an annotation should be added recommending against its use and recommending the use of the sequence. Once the evidence of a ligated Sharada OM is verified, the glyph can/should be replaced.

Conclusion

It is recommended that pending proper evidence of a visually distinct written form for the Bengali OM (or any Telugu/Kannada/other OM), no such character should be encoded in view of maintaining a stable and meaningful policy for OM characters across Indic scripts.

The matter of the Sharada OM may be handled as above.

-o-o-o-