Proposal to encode 1107F BRAHMI NUMBER JOINER (REVISED)

Andrew Glass and Shriramana Sharma

anglass-at-microsoft-dot-com | jamadagni-at-gmail-dot-com

2011-November-2

1. Background

In their Brahmi proposal L2/07-342 in sec 3.5 (pp. 6–7) Baums and Glass noted that in Brahmi:

> *"Multiples of 100 and of 1000 are expressed multiplicatively, with the multiplier following and forming a ligature with 100 or 1000."*

and went on to suggest that:

> *"these ligatures be encoded with ZERO WIDTH JOINER (U+200D)."*

later providing as a sample:

100 4   100-4
(= 104) (= 400)

11063, 11054   11063, 200D, 11054

The Brahmi number forms in this sample come from Bühler 1896 table IX.

In this case the *unligated* form of the symbol of number 100 followed by that of number 4 meaning 100 + 4 = 104 is represented by the sequence BRAHMI NUMBER HUNDRED + BRAHMI NUMBER FOUR whereas the *ligated* form of the same symbol of number 100 followed by that of number 4 meaning 100 × 4 = 400 is represented by the sequence BRAHMI NUMBER HUNDRED + ZWJ + BRAHMI NUMBER FOUR.

Sharma pointed out in a separate communication (L2/10-340) that the suggestion of using ZWJ this way would be against Unicode principles because it makes the joiner semantically significant. Since the joiners are ignored in text search and collation, searches for 104 would turn up 400 and so on which is obviously not desired.

This proposal aims to resolve the issue of representing multiples of hundreds and thousands in Brahmi by encoding a separate script-specific character that will ligate Brahmi numbers.

2. Alternative solutions considered and rejected

Before deciding to encode a script-specific character for this purpose, three alternative solutions were considered and rejected.

2.1. BRAHMI VIRAMA (U+11046)

In general, virama characters are used to suppress an inherent vowel, and may join letter signs to produce conjunct forms. Using BRAHMI VIRAMA with Brahmi numbers as a specialized device would not be in keeping with this principle and would complicate the Virama model unnecessarily. It would also put a burden on users, rendering engines and font developers to manage or mitigate sequences of letters and numbers inappropriately joined by a BRAHMI VIRAMA.

2.2. INVISIBLE TIMES (U+2061)

INVISIBLE TIMES was suggested as a possible solution. However INVISIBLE TIMES is not intended to have a shaping function, but is rather a non-displaying invisible separator. Therefore the use of this code point to form ligatures between Brahmi number signs is not recommended.

2.3. Separately encoded characters for each multiple

Such a solution would entail encoding at least eighteen additional code points and could constrain text encoding if a new multiplier were found. It will be more practical and useful to encode a single code point for this function.

3. New character and its usage

We propose that a new character, BRAHMI NUMBER JOINER, be encoded which would be used in the place of ZWJ in the suggestion of the Brahmi proposal L2/07-342.

1107F   BRAHMI NUMBER JOINER

The code point chosen is 1107F which is empty and comes at the end of the Brahmi block. A similar position at the end of the Tifinagh block has been used for 2D7F TIFINAGH CONSONANT JOINER, which has a consonant joining function in that script.

Now, to indicate 104 by the unligated form the encoded sequence would be:

11064 BRAHMI NUMBER ONE HUNDRED + 11055 BRAHMI NUMBER FOUR

… and to indicate 400 by the ligated form the encoded sequence would be:

11064 BRAHMI NUMBER ONE HUNDRED + *1107F BRAHMI NUMBER JOINER
+ 11055 BRAHMI NUMBER FOUR

Thereby the issue of semantic difference raised by L2/10-340 is addressed.

4. Character name

The proposed name is BRAHMI NUMBER JOINER. This character should be annotated to indicate that it is used to form ligatures between Brahmi numbers (11052–11065) but not Brahmi digits (11066–1106F).

1107F   BNJ   BRAHMI NUMBER JOINER

- Used to form ligatures between Brahmi numbers signifying multiplication

4.1. Visible form

In order to assist text entry and editing, this character should have a visible form which should only be rendered when a number ligature is incomplete, improperly formed (i.e. when the character is used between numbers which would not meaningfully ligate such as BRAHMI NUMBER TWENTY and BRAHMI NUMBER THIRTY or such), or not supported by the selected font. In such cases the control glyph used in the chart should be displayed.

5. Unicode Character Properties

```
1107F;BRAHMI NUMBER JOINER;Mn;9;NSM;;;;;N;;;;;
```

The character properties are the same as those for 11046 BRAHMI VIRAMA. Using the properties for ZWJ was considered but rejected by the UTC on the grounds that doing so would not address the problem that this character seeks to resolve, i.e., to provide a sequence that is semantically distinct. Using ZWJ's properties would result in this character being default ignorable for collation purposes.

The Myanmar block provides a precedent for having two code points with Virama properties in a single code block:

```
1039;MYANMAR SIGN VIRAMA;Mn;9;NSM;;;;;N;;;;;
103A;MYANMAR SIGN ASAT;Mn;9;NSM;;;;;N;;;;;
```

6. Collation

For basic collation the sort weight for the proposed character should be greater than the other numbers and digits already encoded in the Brahmi block (11052–1106F). Ideally, collation mechanisms would be able to interpret number sequences that include Brahmi Number Joiner and so respect its multiplying function.

7. Reference

Bühler, G. 1896. *Indische Palaeographie von circa 350 a. Chr. – circa 1300 p. Chr.* Grundriss der indo-arischen Philologie und Altertumskunde, I. Band, 11. Heft. Strassburg: Verlag von Karl J. Trübner.

8. Official Proposal Summary Form

A. Administrative

1. Title
Proposal to encode 1107F Brahmi Number Joiner
2. Requester's name
Andrew Glass, Shriramana Sharma
3. Requester type (Member body/Liaison/Individual contribution)
Individual
4. Submission date
2011-Oct-13
5. Requester's reference (if applicable)
6. Choose one of the following: This is a complete proposal (or) More information will be provided later
This is a complete proposal.

B. Technical – General

1. Choose one of the following:
1a. This proposal is for a new script (set of characters), Proposed name of script
No
1b. The proposal is for addition of character(s) to an existing block, Name of the existing block
Yes, Brahmi
2. Number of characters in proposal
1 (one)
3. Proposed category
Category B1, specialized small
4. Is a repertoire including character names provided?
Yes
4a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document?
Yes
4b. Are the character shapes attached in a legible form suitable for review?
Yes
5. Fonts related:

a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?

Andrew Glass

b. Identify the party granting a license for use of the font by the editors (include address, e-mail etc.)

Andrew Glass

6a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided?

No.

6b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached?

Yes

7. Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)?

The character is proposed to distinguish the written forms of 104/400 etc in text processes such as searching and sorting.

8. Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script.

See detailed proposal.

C. Technical – Justification

1. Has this proposal for addition of character(s) been submitted before? If YES, explain.

No

2a. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)?

Yes

2b. If YES, with whom?

Stefan Baums, Jost Gippert, Anshuman Pandey, Peter Scharf

2c. If YES, available relevant documents

This proposal has been shared with the above experts

3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included?

Palaeographers and specialists in Brahmi manuscripts and inscriptions.

4a. The context of use for the proposed characters (type of use; common or rare)

Rare

4b. Reference

See detailed proposal.

5a. Are the proposed characters in current use by the user community?

No

5b. If YES, where?

6a. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP?

No

6b. If YES, is a rationale provided?

6c. If YES, reference

7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?

Only one character is proposed. It should located in the same block as the other Brahmi characters.

8a. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?

No

8b. If YES, is a rationale for its inclusion provided?

8c. If YES, reference

9a. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters?

No

9b. If YES, is a rationale for its inclusion provided?

9c. If YES, reference

10a. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character?

No

10b. If YES, is a rationale for its inclusion provided?

10c. If YES, reference

11a. Does the proposal include use of combining characters and/or use of composite sequences?
No

11b. If YES, is a rationale for such use provided?

11c. If YES, reference

11d. Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?

12a. Does the proposal contain characters with any special properties such as control function or similar semantics?
Yes

12b. If YES, describe in detail (include attachment if necessary)
See above, section 3.

13a. Does the proposal contain any Ideographic compatibility character(s)?
No

13b. If YES, is the equivalent corresponding unified ideographic character(s) identified?

13c. If YES, reference:

-o-o-o-