L2/07-238

Towards an Encoding for North Indic Number Forms in the UCS

Anshuman Pandey
University of Michigan
Ann Arbor, Michigan, U.S.A.
pandey@umich.edu

July 31, 2007

1. Introduction

In May 2007, the Unicode Technical Committee (UTC) reviewed a proposal submitted by the present author for encoding "North Indian Accounting Signs" (L2/07-139R) in the Universal Character Set (UCS). The UTC accepted 10 of the 13 proposed characters and tentatively allocated the characters in the Number Forms block (U+2150), as follows:[1]

U+2150 NORTH INDIAN FRACTION ONE QUARTER
U+2151 NORTH INDIAN FRACTION ONE HALF
U+2152 NORTH INDIAN FRACTION THREE QUARTERS
U+2189 NORTH INDIAN FRACTION ONE SIXTEENTH
U+218C NORTH INDIAN FRACTION ONE EIGHTH
U+218B NORTH INDIAN FRACTION THREE SIXTEENTHS
U+218C NORTH INDIAN QUARTER MARK
U+218D NORTH INDIAN PLACEHOLDER MARK
U+218E NORTH INDIAN RUPEE MARK
U+218F NORTH INDIAN WEIGHT MARK

The present proposal is being submitted to the UTC for the purpose of implementing the following recommendations:

1. Change the descriptive name of the characters from NORTH INDIAN to NORTH INDIC.
2. Change the name of NORTH INDIAN WEIGHT MARK to NORTH INDIC QUANTITY MARK.
3. Allocate the characters in a new block to be named "North Indic Number Forms."

2. Change of Descriptive Name of the Characters

The descriptive name NORTH INDIAN for the number forms should be changed to NORTH INDIC for technical reasons. "Indic" is a common term that refers to scripts of the Brahmi family. The number forms and unit marks were used as part of scripts descended from the northern styles of Brahmi. Thus, the term "Indic" more accurately reflects the typological affiliation of the number forms to the northern Indic scripts.

[1] The UTC rejected the three characters NORTH INDIAN INDEPENDENT FRACTION ONE QUARTER, NORTH INDIAN INDEPENDENT FRACTION ONE HALF, and NORTH INDIAN INDEPENDENT FRACTION THREE QUARTERS as it felt that these are composite characters that could be formed from the above accepted fraction signs and already encoded dot characters.

Second, the number forms and unit marks were used as part of scripts, such as Takri and Landa, which are found in regions that lie beyond the boundaries of modern India. The change of name from "Indian" to "Indic" eliminates the possible political constraints presented by the term "Indian."

3. Change of Name of NORTH INDIAN WEIGHT MARK

The name NORTH INDIAN WEIGHT MARK was given in L2/07-139, but was changed to NORTH INDIAN QUANTITY MARK in L2/07-139R. The mark was used for writing weights, measures, and other quantities. The change from WEIGHT MARK to QUANTITY MARK reflects this broader usage. Taking into consideration the recommendation given in Section 2, the new name of the character should be NORTH INDIC QUANTITY MARK.

3.1 New Names for 10 Characters As Allocated

U+2150 NORTH INDIC FRACTION ONE QUARTER
U+2151 NORTH INDIC FRACTION ONE HALF
U+2152 NORTH INDIC FRACTION THREE QUARTERS
U+2189 NORTH INDIC FRACTION ONE SIXTEENTH
U+218A NORTH INDIC FRACTION ONE EIGHTH
U+218B NORTH INDIC FRACTION THREE SIXTEENTHS
U+218C NORTH INDIC QUARTER MARK
U+218D NORTH INDIC PLACEHOLDER MARK
U+218E NORTH INDIC RUPEE MARK
U+218F NORTH INDIC QUANTITY MARK

4. Change of Allocation of Number Forms

Proposal L2/07-139 recommended the allocation of the number forms and unit marks in a new block to be named "North Indian Accounting Signs." The UTC rejected the recommendation for the creation of a new block on concerns of managing the diminishing space in the BMP. Therefore, the characters proposed were tentatively allocated to the Number Forms block (U+2150). The present allocation is not ideal and the characters should be reallocated. There are four options for allocation:

1. Retain the present allocation in the Number Forms block.
2. Allocate within an existing Indic script block.
3. Create a generic "Indic Number Forms" block.
4. Create a "North Indic Number Forms" block.

Based on the discussion of these four options below, it is recommended that a new block be created. The name of the block should be "North Indic Number Forms" and the accepted number forms and unit marks should be allocated within it.

4.1 Retain the present allocation

The allocation of the north Indic number forms to the Number Forms block is suitable only on the bases of semantics and managing empty space in the BMP.

The disadvantage of using the Number Forms block is that it is inherently associated with the Latin script. The block contains Roman numerals and vulgar fractions. Allocating the north Indic forms here is

semantically appropriate, but typologically disparate. As such, users will not expect to find Indic forms in the Number Forms block. Moreover, the RUPEE MARK and QUANTITY MARK are not semantically aligned with the Latin forms.

Another disadvantage of placing the north Indic number forms in the Number Forms block is diminishing space within the block itself. As indicated in the Unicode Pipeline, four new Latin forms were allocated to the block in November 2006.[2] Given such activity, the remaining space within the block should be reserved for more suitable Latin forms. If the Indic number forms are placed within the Number Forms block and additional Latin forms are proposed in the future, then the more germane Latin forms would need to be placed elsewhere in the BMP. Such a situation might contradict the space-management concerns that initially governed the UTC's allocation of north Indic forms to Number Forms.

4.2 Allocate within an existing Indic script block

Since the proposed number forms belong to the Indic family of script, it might be appropriate to allocate the characters within existing Indic script blocks. Such an action is more appropriate than placing Indic characters in a block intended for Latin-based characters. However, adding the number forms to an existing script has negative consequences. Encoding the characters in an existing Indic block creates the impression that the number forms belong to that particular script. This is problematic since that number forms and unit marks are intended for general use across north Indic scripts. If the characters are placed within an existing block, it would be necessary to instruct users and implementers that the number forms are to be used with other scripts, not solely with the given block. It would also be necessary to explain that the number forms must not be added solely to fonts of the particular script block, but may be added to other north Indic fonts as well.

4.3 Create a generic "Indic Number Forms" block

Just as the Number Forms block at U+2150 was created to house such forms as found in Latin, a new block should be created for Indic scripts for the same purpose. The north Indic number forms and unit marks could be allocated in a block named "Indic Number Forms," which in principle would be congruent with existing supplemental blocks for script families. An "Indic Number Forms" block would contain not only the north Indic number forms, but could be used to encode fraction signs and unit marks from all other Indic scripts. Number forms of Malayalam and other southern Indic scripts could conceivably be encoded together with northern Indic ones on grounds of semantic similarity. Such a pan-Indic number forms block could potentially eliminate the problem of dealing with allocations for supplemental blocks for Indic scripts. Theoretically, this block could serve as a space to unify number forms that are already encoded as part of the repertoires of individual Indic scripts for generic use across script sub-families; for example, some Bengali currency enumerators are also used in the Assamese, Maithili, and Oriya scripts.

A similar recommendation was made in a proposal submitted to the UTC by N. Ganesan for a "South Indian Supplement" block. The proposal sought to encode Malayalam fractions and letter-numerals. Rather than place the proposed characters within the Malayalam block, the proposal recommended the creation of a generic block that would serve as a container for various number forms found in southern Indic scripts. However, as indicated in the Unicode Pipeline, the UTC allocated the Malayalam fractions as part of the Malayalam block.

The disadvantage of a generic Indic Number Forms block is that while there may be wide semantic similarity, the scope of such a block is too large. Just as script-specific fraction signs and unit marks are

[2] http://unicode.org/alloc/Pipeline.html

encoded as part of the character repertoire of specific script, fraction signs and unit marks that belong to a range of scripts should be encoded in a manner that they may be used with all of the relevant scripts.

4.4 Create a "North Indic Number Forms" block

The north Indic number forms and unit marks constitute a specialized set of characters that belong to a numeric notation system that was used throughout northern India and Pakistan. Unlike the fraction signs, weight marks, and other unit marks recently allocated within the Malayalam and Telugu blocks, the north Indic number forms are not associated with any specific script. Rather, the north Indic number forms and unit marks were added as part of the repertoire of various scripts, and the standard shape of the forms was maintained. Given these characteristics, the forms should be not be allocated to an existing script or number forms block, but should be given an independent allocation.

A separate "North Indic Number Forms" block would require less than one column. There is one column between the Syloti Nagari and Phags-Pa blocks at the range U+A830...U+A83F. Placing the number forms at this range would help the UTC to patch up the BMP. These 16 code-points are sufficient for encoding the 10 accepted characters and would leave 6 unassigned positions for the possible addition of signs in the future. One drawback to this recommendation is that implementation and management of the BMP could be complicated by the creation of single column blocks. However, the "North Indic Number Forms" block should be considered a complete set of characters. Also, it is highly unlikely that the repertoire of the block would grow beyond 16 characters.

The size of the column at U+A830...U+A83F limits the placement of complete scripts or specialized sets of characters at this range in the future. However, another drawback is that adjacent to this column are four unassigned code-points (U+A82C...U+A82F) that belong to the Syloti Nagri block. The possibility of 4 new additions to Syloti Nagri would fill the block and any additions beyond these 4 characters would force the non-contiguous allocation of future Syloti Nagri characters.

It is not necessary that the number forms and unit marks be encoded in the BMP. There is sufficient space in the SMP or another plane to encode these characters in an independent block.

The benefits of encoding these characters in an independent block, despite the minimal size of the block, outweigh the disadvantages of doing so. An independent allocation would assist in establishing the number forms and unit signs as an independent set of specialized characters. A separate "North Indic Number Forms" block would make it clear that the characters are to be used as a supplement to the various north Indic scripts. Moreover, an independent allocation would facilitate the identification of the characters for user and implementers.

The unified encoding of the north Indic number forms and unit marks within an independent "North Indic Number Forms" block – instead of within existing Indic script or number form blocks – will enable their use across north Indic writing systems in a manner that reflects historical and contemporary practices.

5. Conclusion

The names of the characters accepted for encoding should be

U+A830 NORTH INDIC FRACTION ONE QUARTER
U+A831 NORTH INDIC FRACTION ONE HALF
U+A832 NORTH INDIC FRACTION THREE QUARTERS

U+A833 NORTH INDIC FRACTION ONE SIXTEENTH
U+A834 NORTH INDIC FRACTION ONE EIGHTH
U+A835 NORTH INDIC FRACTION THREE SIXTEENTHS
U+A836 NORTH INDIC QUARTER MARK
U+A837 NORTH INDIC PLACEHOLDER MARK
U+A838 NORTH INDIC RUPEE MARK
U+A839 NORTH INDIC QUANTITY MARK

and the characters should be allocated to a new block to be named “North Indic Number Forms.”