PROPOSALS FOR MALAYALAM AND TAMIL SCRIPTS' ROOT ZONE LABEL GENERATION RULES

Publication Date:	23 November 2018
Prepared By:	IDN Program, ICANN Org

Public Comment Proceeding	
Open Date:	25 September 2018
Close Date:	7 November 2018
Staff Report Due Date:	21 November 2018

Important Information Links
Announcement
Public Comment Proceeding
View Comments Submitted

Staff Contact:	Sarmad Hussain	**Email:**	sarmad.hussain@icann.org

Section I: General Overview and Next Steps

The Neo-Brahmi Script Generation Panel (NBGP) was formed by nine communities that use scripts derived from the Brahmi script. NBGP is developing Root Zone Label Generation Rules (LGR) for Bengali, Devanagari, Gujarati, Gurmukhi, Kannada, Malayalam, Oriya, Tamil and Telugu scripts. The GP has published the proposals for eight LGRs from these nine scripts in three sets, releasing proposals for the scripts which share cross-script variant code points together to the extent possible. The first and second sets have undergone public comment and this third set included the following proposals: (1) *Proposal for the Malayalam Script Label Generation Rules for the Root Zone,* (2) *Proposal for the Tamil Script Label Generation Rules for the Root Zone.* As per the LGR Procedure, these proposals were posted for public comment to allow those who have not participated in the NBGP to make their views known. Based on the feedback, the NBGP will finalize each proposal for its evaluation and integration into the Label Generation Rules for the Root Zone.

Section II: Contributors

At the time this report was prepared, a total of ten (10) community submissions had been posted to the forum. The contributors, both individuals and organizations/groups, are listed below in chronological order by posting date with initials noted. To the extent that quotations are used in the foregoing narrative (Section III), such citations will reference the contributor's initials.

Organizations and Groups:

Name	**Submitted by**	**Initials**
Myanmar Script Generation Panel	Thin Zar Phyo	MMGP
Wikimedia Foundation Language Engineering	Santhosh Thottingal	WFLE
International Centre for Free and Open Source Software	Dinesh Lal D L	ICFOSS

Individuals:

Name	**Affiliation (if provided)**	**Initials**
Liang Hai		LH
Cibu		CI

Name	Affiliation (if provided)	Initials
Gowtham Raghunathan		GR
R Selvaraj		RS
Ajay		AJ

Section III: Summary of Comments

<u>General Disclaimer</u>: This section intends to summarize broadly and comprehensively the comments submitted to this public comment proceeding but does not address every specific position stated by each contributor. The preparer recommends that readers interested in specific aspects of any of the summarized comments, or the full context of others, refer directly to the specific contributions at the link referenced above (View Comments Submitted).

MMGP reviewed the the NBGP LGRs and makes following comments:

MMGP1. MMGP notes that in Myanmar LGR proposal, the following sets are being considered as variant code points.

- Set1: ဂ(U+1002) MYANMAR LETTER GA and റ (U+0D31) MALAYALAM LETTER RRA
- Set2: ဝ (U+101D) MYANMAR LETTER WA and ഠ (U+0D20) MALAYALAM LETTER TTHA

MMGP encourages NBGP to review these variant code points in Malayalam LGR proposal.

MMGP2. MMGP shares a list of Malayalam-Myanmar confusable code points being considered by MMGP and comments that it is useful for readers if the both Malayalam and Myanmar LGR proposals share the same list.

WFLE reviewed the Malayalam proposal and makes the following comments:

WFLE1. The case of റ്റ is similar to the case of ള്ള in the document. A font that does not stack the റ +് + റ can render it in horizontal format. So a word like മാററ് can be spoofed by applying virama to the last two റ. It is better to add a WLE rule similar to ള്ള for this case as well.

WFLE2. The document should not conflict with Unicode version 11, Chapter 12. In the table under 6.1, 1a, 1b, 1c – all three versions should be allowed as variants and the LGR proposalshould not block any of them.

ICFOSS reviewd the Malayalam proposal and makes the following comments:

ICFOSS1. ICFOSS suggests including '_' (underscore) / '-' (hyphen) due to the absence of Zero Width Non-Joiner.

ICFOSS2. ICFOSS suggests the following pairs are in-script variants code point sequences.
- റ് +റ (0D31 + 0D4D + 0D31) -->ററ / റ+റ(0D31 + 0D31)--->ററ
- ള് +ള(0D33 + 0D4D + 0D33) -->ളള / ള+ള(0D33 + 0D33)--->ളള

ICFOSS3. ICFOSS comments that the transliteration of acronyms creates some issues. For example:

- HDFC ---> എച് ഡി എഫ് സി
  ---> എച്ച് ഡി എഫ് സി
- tata ---> ററാററ
  ---> റ്റാറ്റ

LH reviewed the Tamil proposal and makes the following comments:

LH1. LH suggests that the following points or sections should be revised.

(1) Throughout the document, the ISO15959 transliteration should be consistently used

(2) In section 3.1, Figure 1, the name and description below are mismatched

(3) In section 3.1.1, the distinction between Tamil traditional grammar and Unicode terminologies should be clarified

(4) In section 3.3.2, the examples should be in a clear format

(5) In section 4.1.2.4, LH suggests using the rationale “that is in Unicode Normalization Form C (NFC)” instead because of its rare usage

(6) In section 5.2, the text points to the empty reference in section 3.2

(7) In section 5.2, Table 5, it should be noted that the “Indic syllabic category” column is not about the Unicode character property of the same name

(8) In section 5.5, LH suggests using the word “notation” instead of “variables”

(9) In section 5.5.4, LH suggests that it does not need to follow the Devanagari practice and its Akshar formation can simply be explained with a single base consonant. The special case of śrī and kṣV should be discussed as the exceptional cases

(10) In section 6.4, Table 21, there is an incorrect rendering of the Malayalam text

LH2. LH raises the following questions and discussion points.

(1) In section 3.3.4, the restriction of <…, visarga, visarga, …> conflicts with the WLE rules in section 7

(2) In section 5.2.1, Table 6a, is the “=” in the second column intentional?

(3) The case section 6.3, should already be eliminated by the IDNA 2008

(4) In appendix A, Table 22, based on the same level of similarity, the following pairs (and probably more) should also be considered:

- U+0B89 TAMIL LETTER U and U+0D09 MALAYALAM LETTER U
- U+0BB5 TAMIL LETTER VA and U+0D35 MALAYALAM LETTER VA
- U+0BB7 TAMIL LETTER SSA and U+0D37 MALAYALAM LETTER SSA

LH3. LH suggests the pattern representation of Malayalam labels as:
`C[M][B|X] | V[B|X] | C[U+0D41]H | L`

LH4. In section 6.1.3, LH agrees with allowing the applicant to make both encodings aliases to each other.

CI reviewed the Malayalam proposal and makes the following comments:

CI1. In section 6.1, it is proposed to disallow <chillu-n, virama, rra>. However, this conflicts with Unicode Standard Version 11.0.0 (§12.9 page 506 table12-38) where <chillu-n, virama, rra> is the prescribed sequence for the form {chillu-n base, rra below-base}. Therefore, the sequence <chillu-n, virama, rra> should be allowed.

GR reviewed the NBGP proposals and makes the following comments:

GR1. GR comments that English is a language which has unique letters and unique pronounciation. Other languages have identical letters and identical pronounciation which will facilitate cyber thefts.

RS reviewed the Malayalam proposals and makes the following comments:

RS1. When typing with Unicode input tools another font issue encountered is that ററ is rendered as റ്റ in some cases.

RS2. More explanation on the similarity case between Malayalam anusvara " ം" and English "o" should be provided.

AJ comments that this is very useful for a lot of people in the state of Kerala.

LH reviewed the Malayalam proposal and makes the following comments:

LH1. LH suggests that the following points or sections should be revised.

(1) In section 3.1–3.3, the script's history should be removed or moved to an appendix

(2) In section 3.6, the width of Table 5 should be adjusted to avoid line breaks

(3) In section 3.6, the text regarding ZWJ in multiple places should be revised

(4) In section 3.6,tThe text "ICANN's Maximal Starting Repertoire (MSR) for IDN LGR is based on these exclusion rules for ZWJ and ZWNJ." should be revised or rule specified

(5) In section 6.1, the analysis should be revised:
   - In case 1a, NBGP should work with Unicode Consortium and ensure the consistent recommendations on the <chillu n base, below-base rra sign> encoding
   - In case 1b, both <NA, VIRAMA, ZWJ, RRA> and <CHILLU N, VIRAMA, RRA> should be mentioned
   - More explanation is needed for disallowing <CHILLU N, VIRAMA, RRA> while allowing <NA, VIRAMA, RRA>, when both sequences have rendering problems

- More explanation is needed for mapping <chillu n base, rra base> as blocked variant of <chillu n base, below-base rra sign>. Ordinary fonts should be able to handle and won’t cause confusion

(6) In section 3.7, LH suggests various revisions detailed in comment submitted

(7) In section 6.2.1, Table 10 has a rendering issue of Tamil glyph in the set 6. The last column should be rendered without dotted circles

(8) In section 10, the Unicode Standards does not recommend “not to use the sequence[s]”

LH2. LH raises the following questions and discussion points.

(5) In section 3.5, Sanskrit using Malayalam script should have its own EGIDS rating for such an evaluation

(6) In section 3.7, it is unclear why the consonant letter ള ḷa is missing

(7) LH agrees with discussing ള്ള vs ളള pair. However he comments that the LGR is over restricted and more research is needed for the following cases:

- combinations when inter-word spaces are removed from a sequence of words as it can introduce many more sequences that were previously considered highly limited, e.g. a much larger number of ളള
- combinations with the final glyph sequence including the reordered glyphs, such as pre-base vowel signs, which can break an otherwise confusable sequence, eg, ളള + ◌െ → ളളെ

(8) In section 7, rules 5 and 6 seems to be limited by phonology and spelling conventions

(9) In section 7, rule 7 doesn’t seem to be consistent with the restrictions suggested in §6.1

(10) For section 7, LH notes that in the Unicode Standard’s Core Specification suggests (see Table 12-33, page 504, in the referred Core Spec 10.0), a samvruthokaram does not only appear at the end of a word, but it can also appear as an independent vowel letter (typically a word-initial structure) or be followed by a anusvaram. The inconsistency between the Unicode’s claim and this document’s analysis must be addressed, and the WLE rules might need to be less restrictive

LH3. LH suggests the pattern representation of Malayalam labels as:
`C[M][B|X] | V[B|X] | C[U+0D41]H | L`

Section IV: Analysis of Comments

<u>General Disclaimer</u>: This section intends to provide an analysis and evaluation of the comments submitted along with explanations regarding the basis for any recommendations provided within the analysis.

These comments are being submitted to the Neo-Brahmi Generation Panel for their consideration and incorporation (as required) in the final version of the proposals.