

update from
APTLD Working Group on

Linguistic Guidelines for Domain Registrations in Indic Languages

Gihan Dias

.LK

Guidelines for Domain Registrations

- For each TLD or 2LD, we need to decide:
- Scripts which are registered in that domain
- Allowed characters within the script
 - disallowed characters in the script (e.g. numerals)
- defined in IDN Tables

“ICANN strongly recommends collaboration in the development of the IDN Tables”

Indic Scripts

- Scripts used in South Asia
 - Devanagari, Bengali, Telugu, Tamil, ...
 - Burmese, Sinhala
 - RTL scripts e.g. Urdu
- Follow a common Pattern
 - Basically use the same letters
- Need common guidelines + specific rules for each script

Issues with Indic Scripts

- “Letters” often comprise 2 or more Unicode characters
- Issues are often with combinations of characters, not just individual characters
- Allowed character sequences
 - modifiers
 - joiners
- Disallowed character sequences
- Inter-character dependencies

Proposal

- Form Working Group to form guidelines
 - Hosted by APTLD
 - open membership
- Objective – publish RFC
- Will be used by
 - TLDs and 2LDs
 - other registries (e.g. corporate)
 - software developers
- Timeline – 6 months?

Working Group

Han Chuan, Lee

Muthu Nedumaran

Ram Mohan

Azmah Abd Malik

Noor Harnidda Khalid

Dr. Govind

Swaran Lata

Tulika Pandey

sandoche BALAKRICHENAN

Balachandaran, G

Sarveswaran, K

Chamara Disanayake

Gihan Dias

Narayanan A

Mani Manivannan

Methodology

- Tamil is a multi-national language
 - used in Sri Lanka, Malaysia and Singapore in addition to India
 - இலங்கை and சிங்கப்பூர் have already passed the IDN ccTLD string evaluation stage
- Decided to start with Tamil and then move on to other languages
 - Tamil is in many respects simpler than other Indic languages

Methodology

- Review proposed Tamil IDN tables from Malaysia, Singapore, India and Sri Lanka
- Build a single Harmonised table

Tamil IDN Table

Tamil_Aytham = U+0B83 # (ஃ)

Tamil_Letter_A = U+0B85 # (அ)

Tamil_Letter_AA = U+0B86 # (ஆ)

Tamil_Letter_I = U+0B87 # (இ)

Tamil_Letter_II = U+0B88 # (ஈ)

Tamil_Letter_U = U+0B89 # (உ)

Tamil_Letter_UU = U+0B8A # (ஊ)

Tamil_Letter_E = U+0B8E # (எ)

Tamil_Letter_EE = U+0B8F # (ஏ)

Tamil_Letter_AI = U+0B90 # (ஐ)

Tamil_Letter_O = U+0B92 # (ஒ)

Tamil_Letter_OO = U+0B93 # (ஔ)

Tamil_Letter_AU = U+0B94 # (ஔ)

Tamil_Letter_KA = U+0B95 # (க)

Tamil_Letter_NGA = U+0B99 # (ங)

Tamil_Letter_CA = U+0B9A # (ச)

Tamil_Letter_JA = U+0B9C # (ஜ)

Tamil_Letter_NYA = U+0B9E # (ஞ)

Tamil_Letter_TTA = U+0B9F # (ட)

Tamil_Letter_NNA = U+0BA3 # (ண)

Tamil_Letter_TA = U+0BA4 # (த)

Tamil_Letter_NA = U+0BA8 # (ந)

Tamil_Letter_NNNA = U+0BA9 # (ன)

Tamil_Letter_PA = U+0BAA # (ப)

Tamil_Letter_MA = U+0BAE # (ம)

Tamil_Letter_YA = U+0BAF # (ய)

Tamil_Letter_RA = U+0BB0 # (ர)

Tamil_Letter_RRA = U+0BB1 # (ற)

Tamil_Letter_LA = U+0BB2 # (ல)

Tamil_Letter_LLA = U+0BB3 # (ள)

Tamil_Letter_LLLA = U+0BB4 # (ழ)

Tamil_Letter_VA = U+0BB5 # (வ)

Tamil_Letter_SHA = U+0BB6 # (ழ)

Tamil_Letter_SSA = U+0BB7 # (ஶ)

Tamil_Letter_SA = U+0BB8 # (ஷ)

Tamil_Letter_HA = U+0BB9 # (ஹ)

Tamil_Vowel_Sign_AA = U+0BBE # (ஃ)

Tamil_Vowel_Sign_I = U+0BBF # (ஐ)

Tamil_Vowel_Sign_II = U+0BC0 # (ஊ)

Tamil_Vowel_Sign_U = U+0BC1 # (஋)

Tamil_Vowel_Sign_UU = U+0BC2 # (஌)

Tamil_Vowel_Sign_E = U+0BC6 # (எ)

Tamil_Vowel_Sign_EE = U+0BC7 # (ஏ)

Tamil_Vowel_Sign_AI = U+0BC8 # (ஐ)

Tamil_Vowel_Sign_O = U+0BCA # (ஔ)

Tamil_Vowel_Sign_OO = U+0BCB # (க)

Tamil_Vowel_Sign_AU = U+0BCC # (஖)

Tamil_Sign_Pullu = U+0BCD # (ட) (Unicode Name=TAMIL SIGN VIRAMA)

ZWNJ = U+200C

Hyphen-Minus = U+002D

Tamil String Sequences

In many cases issues are with sequences of characters

- allow ik + ssa as either a single glyph (சுஷ) or separate glyphs (க்ஷ)
- disallow old form of Shri (ஸ+ஃ+ர+ஃ)
- disallow a LLA after a consonant with a Kombu (e.g. கௌ) unless it is modified by a vowel sign or Pulli to avoid confusion with TamilConsonant +Vowel Sign AU (கௌ)

Tamil String Sequences (cont.)

- disallow a LLA after Letter O (ஓ ள) unless it is modified by a vowel sign or Pulli to avoid confusion with Letter AU (ஔ)
- a vowel sign or a pulli (virama) can only follow a consonant and is not valid elsewhere
- allow consonants, vowels, Aytham, European numerals anywhere (unless disallowed by previous rules)
- anything else is invalid

Next Steps

- Reach consensus on Tamil
- Extend to other Indic Scripts
- Multi-national Scripts
 - Devanagari - India, Nepal Pakistan
 - Bangla – India Bangladesh
- National Scripts
 - India, Sri Lanka

Next Steps (cont.)

- Expand the Working Group
- Need members from Bangladesh, Nepal, Pakistan, etc.

Objective

- Publish an informational RFC on IDN Registrations in Indic Languages