CDAC-IDN-Critique

Malayalam IDN Policy Draft by CDAC - Critique by SMC

Swathanthra Malayalam Computing Project http://smc.org.in

SMC's Response to CDAC's Answers to the Critique (dated 2/12/2010)

We are not satisfied with the answers given to our questions. The response seems to be prepared by somebody who does not know the scope of IDN. The replies are based on things that are not at all related to IDN. We have tried to explain why your answers are completely wrong. See the responses given below. CDAC seems to be confused about the Malayalam orthography, Unicode standard, Font , rendering etc.

One of the main thing that we want to emphasis is, standards should not be drafted based on the existing applications. The response from CDAC says, it did extensive study on existing browsers and their address bar behavior. Even one variant table entry is associated with Microsoft Internet Explores buggy behavior. Sorry to say that this is ridiculous. Standards and Policies are for the implementations in the future. It is the guidelines for the applications to be developed in future. It is not a study of existing application behavior. This is a trivial fact any person with minimal technical qualification is aware of. CDAC seems to be missing this point.

Orthography is style of writing. It is a users choice for visualizing the content. The content remains same whether one use traditional orthography or new orthography. What IDN policy need to take care of is avoiding all cases of spoofing. That should be independent of orthography. IDN can not assume , or cannot enforce that a user should be using only a particular font with Internet browsers. The CDAC's draft policy makes such a false assumption. It states that Malayalam IDN is not designed for Traditional orthography. This argument comes from ignorance about the basic concepts of Language technology. Unicode does not define anything about Malayalam new and old orthography. It is all about users choice. When CDAC's policy states "Modern Orthography", it should define what it is. It is not all defined technically or even in usage pattern. It is just a user's convenience.

We demand a process to re-draft this policy document. This is the age of open standards. CDAC should take initiatives to consult with all stakeholders- language experts, language computing experts, developers.

Criticism on the policy

General Comments

1. The variant table is defined based on random glyphs taken from a list of 900+ possible glyphs for Malayalam. No explanation is given on how two entries in variant table become homo morphs. One entry in variant table is just because of the fact that one is mirror image of other. Since b, d are not excluded from English, there's no need to exclude mirror imaged glyphs in variant table.

CDAC's Response: The IDN system devised for Malayalam is based only on the modern script. It doesn't address the old script or the fonts based on old script. Also, a detailed study was done before proposing homographs in each of the languages. The study included observing the visual form of the conjunct in the point size of the Address bars of major browsers. The mirror imaged nature of the glyphs was not the criterion for the two glyphs to be qualified as variants.

> Since CDAC took that much effort to check visual spoofing, i just like to know the basic glyph set in modern orthography of Malayalam --Jinesh.k 06:26, 4 December 2010 (PST)
>
> Please understand that IDN has no relation with the orthography. Orthography is kind of writing style and decided by user's choice on fonts. It is at higher level and close to user, while IDN has nothing to do about the choice of fonts by users. The distinction between Modern Orthography and Old orthography is not technically defined. Unicode or Unicode capable applications never bother about whether user use an Old lipi font or Traditional font. CDAC or IDN policy cannot say that user should use only a certain set of fonts to work with IDN. If CDAC says IDN does not support traditional orthography, it does not make any sense at all for some one who know what is Unicode, Orthography and IDN. - സന്തോഷ് 23:53, 4 December 2010 (PST)
>
> Forcing a user to chose a particular orthography font based on bugs in one browser is not acceptable. IDN standard should not ban most popular fonts currently in use, majority of digital Malayalam users are happy with traditional fonts and there is no reason to stop it. CDAC cannot and should not decide what fonts a user choses in his/her computer. IDN should work with any Malayalam font, which complies with Unicode version 5.0 and correctly implement Malayalam language rules for conjunct formation. Unicode 5.1 is controversial and many issues are not answered and it would be a tragedy for Malayalam language if we decide to follow unicode 5.1 blindly without considering its impact on Malayalam language for years to come. Pravs 06:30, 7 December 2010 (PST)
>
> If mirror image was not a criteria for variant table, explain how ത്സ/ത്ഡ , സ്സ/ഡ്ഡ qualify to the variant table? - സന്തോഷ് 00:00, 5 December 2010 (PST)
>
> Why browser behavior is studied for drafting IDN standard? Does it mean that if this policy was drafted 10 years, back, none of the Malayalam characters will be allowed in IDN? 10 years back, Malayalam rendering was pathetic in browsers. Or the current policy is going to change when browsers improve the rendering and their address bar behavior later? How can a standard drafted based on its Implementation? - സന്തോഷ് 00:00, 5 December 2010 (PST)
>
> A standard should not be based on buggy implementation of Malayalam rendering, that is encouraging inability. Any wrong implementation should be corrected. Malayalam language in digital domain should not be kept at the mercy of some corporations who does not care about the language. Pravs 06:34, 7 December 2010 (PST)

1. Visually identical glyphs are not the only entries to be considered for the variant table. Unicode chart itself has ambiguous dual representations for the same code point without canonical equivalence. An example for this is au signs in Tamil and Malayalam. ◌ൗ-◌ൗand ◌ൗ - ◌ൗ . The document does not consider these special cases.

CDAC's Response: The IDN policy does not permit the entry of syllables having structure CMM or MCM, where M stands for Matra or vowel sign. The ABNF rules takes care of this.

That is wrong. ◌ൗ - ക◌ൗ are neither CMM nor MCM case. It is single code pointed Mathra(vowel signs), appearing with consonants in CM format alone. -സന്തോഷ് 01:04, 5 December 2010 (PST)

2. There are different orthographic forms for many glyphs in Malayalam. The variant table does not address different scenarios arising while considering the visual similarity. For example in traditional orthography TTA is written in stacked form. While in modern orthography it can be written in non-stacked form and this non-stacked form is visually identical to two RA sequence (ൎൎ).

CDAC's Response : Only the stacked form is considered to be the conjunct TTA in modern orthography.

i don't really understand the logic. A normal user is easily spoofed with �gൎ and ൎൎ. If we go by CDAC's logic another inorganic standardisation will be the result --Jinesh.k 06:26, 4 December 2010 (PST)
As noted before, please don't answer questions about IDN using Modern/Old orthography distinction. Nobody with proper Malayalam knowledge will say that nonstacked TTA is not TTA. The nonstacked form is explained in detail in Unicode 5.1.0 standard. see http://unicode.org/versions/Unicode5.1.0/ CDACs statement is a contradiction to this.

ABNF rules

1. Section 2 says ക് as pure consonant of ക. Chillu of ൿ is considered as pure consonant of ka.

CDAC's Response: The policy document doesn't address the obsolete characters in the script, although those characters might have been included in Unicode code chart.

On what basis CDAC decided that a character is obsolete? Does CDAC understand that people write the name "CDAC" in Malayalam using Chillu of K like സിഡാൿ ? -സന്തോഷ് 00:20, 5 December 2010 (PST)

2. Section 2.a says CM can be followed by only D (anuswara) or X (visarga). This excludes the Samvruthokarams of Malayalam. All consonant can have cons + u vowel sign + virama and forming samvruthokaram form of that consonant. Examples: ള്, ങ്, പ്, ത്.

CDAC's Response: The use of samvruthokarams is considered to be the part of traditional orthography which the policy doesn't permit.

We already explained that orthography is not even a subject of discussion here. Can CDAC provide a definition on what is traditional orthography and what is modern orthography? With a list of "allowed" "characters" in modern orthography which is present in traditional Orthography? What is meant by "permitting"? What stops from a user to use a so called modern orthography font to write samvruthokaram? - സന്തോഷ് 00:20, 5 December 2010 (PST)
It is completely unacceptable for CDAC to dictate choice of orthography for any user. It does not make any sense as well. Samvruthokaram can be written using modern orthography as well as traditional orthographies. Pravs 06:40, 7 December 2010 (PST)

1. Section 3.a restrict the count of consonant in syllable as 4. But ശ്ര്യ has 5 consonants

   CDAC's Response : Complex conjuncts like ശ്ര്യ have been simplified in modern orthography.

   Just wanted to know how a conjunct can be simplified? If "simplified" will the number of consonants reduce? Will the number of Unicode characters reduce?! CDAC should understand the difference between a conjunct and glyph before answering the question. -- സന്തോഷ് 00:20, 5 December 2010 (PST)
2. Section 3.b excludes syllables with samvruthokram like ക്കു് .
3. Section 4 states a chillu can be followed by a vowel sign. Since chillu is dead consonant, there is no possibility of having virama after chillu.

   CDAC's Response : The document doesn't state that a chillu can be followed by a vowel sign. The observation that a virama can appear after a chillu is based on the recommendation of Unicode
4. The example used for LHC - ൻ്റ does not exist in printing or digital format. None of the input methods or Malayalam writers ൻ്റ in this way. The sequence for nta is ന + ് + റ . ie there is no LHC sequence in Malayalam.

   CDAC's Response: The ൻ്റ happened because none of the rendering engines available today does support the rendering of 'nTa' in the Unicode 5.1 way. as displayed in the document is the wrongly rendered form of the conjunct 'nTa'. This 5.1 official document on rendering the conjunct 'nTa'.

   CDAC said, it did extensive study on rendering in various browser address bars and took policy decision was based on that. So for this, no study was conducted? Never noticed that browsers are not able to render this properly? -സന്തോഷ് 00:57, 5 December 2010 (PST)
   It seems, no systematic process or criteria were used to arrive at the conclusions. Pravs 06:47, 7 December 2010 (PST)
5. Since LHC is invalid for Malayalam, including L = ൻ , section 5 of the document cancels itself.
6. Because of argument #6, section 6 also cancels itself.
7. Because of arguments #1 to #8 the IDN rule "Consonant Sequence → *3(CH) C [H| D|X|M[D|X]] | L[HC[D|H|M[D]]]" is completely wrong and need to be reformulated.

Restriction Rules

1. Section 2 says "H is not permitted after V, D, X, M, digit and dash" This is wrong since samvruthokaram requires H after V

   CDAC's Response : See the explanation for section 2 under ABNF Rules
2. Section 7 says H can follow L if it is followed by റ , This is wrong as explained above. L can never followed by H. It can only followed by C

   CDAC's Response : See the explanation for section 5 under ABNF Rules

nta criticism

1. This document does not address the case of stacked and non stacked forms of nta, which are interchangeably used. For example എന്റെ can be spoofed with എൻ്റ. Severity of this issue is increased by having one more sequence to represent the same conjunct (ൻ + ് + റ) is introduced in Unicode 5.1

    CDAC's Response : Ans : Modern orthography treats ൻറ as 'nRa' and ൻറ as 'nTa'. The interchangeable usage of stacked and non-stacked forms for the conjunct 'nTa' is wrong by convention.

    One of the main motivations for IDN restrictions is to avoid spoofing, and this policy does not address legitimate case of multiple encoding for 'nTa'. This is a glowing example of complete disregard for the language and obedience to faulty implementation by a big corporation. Though it cannot be solved at IDN level (it is an issue at encoding level), there should be an effort to address severity of this issue. To me, the only solution seems to be dumping the 5.1 version and Microsoft's version of nTa in favour of Malayalam nTa, as per laguage rules. Pravs 06:58, 7 December 2010 (PST)

Chart of allowed characters

1. Malayalam chillus - the 5.1 version ൿ is removed from the tables. which is having same characteristics and use cases of other chillus. So excluding it from the allowed code points does not make any sense. Moreover the existing chillu representation - non-atomic - is not mentioned in the document at all.
2. Malayalam au sign - ൊ is not allowed. Instead the au length mark ൗ is provided. The inscript standard does not allow one to type ൗ and allows only ൊ. Other input methods allows to type both. But the document does not say anything on the equivalence of both. Allowing both vowel signs is also a spoofing issue. And hence this should be handled in variant table.

    CDAC's Response : The inscript standard being revised. The new standard allows both the characters to be inputted. For restricting spoofing and phishing, only one form i.e. ൗ by IDN policy (used in modern text) has been allowed

    This also shows CDAC is completely clueless about the difference between character encoding and input methods. How is CDAC going to prevent me from using ൊ with old inscript standard or entering the unicode value directly? This is an encoding issue and has to be solved at encoding level by providing equivalance. Trying to fix encoding issue by mandating a specific input method is like prescribing ointment when you need a surgery (ചൂക്കുവെള്ളം കുടിച്ചാൽ കാൻസർ ഭേദമാകും എന്ന് പ്രതീക്ഷിയ്ക്കുന്ന പോലെയാണത്.) Pravs 07:08, 7 December 2010 (PST)
    Our response to the CDACs revised Inscript standard- CDAC-Inscript-Critique - സന്തോഷ് 07:10, 7 December 2010 (PST)

Variant Table and Visual Spoofing

Variant table is not logical. Only ള and ള makes sense. None of the other entries should be considered as spoofing. ന and ന is not even close. Mirror images are already used in Latin, eg. b

and d. Hence ന്ന and ണ്ണ cannot be blocked. Moreover it is not clear why the same logic does not apply for ന and ണ. It did not consider the case of ന്ന and non stacked form of ൃ common in new lipi.

> CDAC's Response : The variant table is based on the observations how Malayalam characters and conjuncts are rendered in the address bars of standard browsers like IE, Mozilla and Safari. While ന and ന are perfectly rendered in Mozilla and Safari, they are not legibly rendered in various versions of IE. The mirror imaged nature of the glyphs was not the criterion for the two glyphs to be qualified as variants. Also note that the variant table is not a full-proof mechanism which can prevent spoofing.
>
> > We cannot keep our language hostage to faulty software from one company. Bugs needs to be fixed, not the other way around ie, standards are not drafted based on buggy software. It is completly unacceptable, CDAC should ask Microsoft to fix its rendering in IE. Pravs 07:15, 7 December 2010 (PST)

Even though similarity is considered, dual encoding is not mentioned. In case of dual encoding of chillus, both forms (atomic chillu and consonant chandrakkala ZWJ) of chillus will look SAME.

> CDAC's Response : IDN policy doesn't allow control characters such as ZWJ and ZWNJ to be part of domain names.
>
> > ZWJ and ZWNJ are part of unicode standard and is required for Malayalam (even though atomic chillus might solve ZWJ issue, there is no substitute for ZWNJ). We need these characters for using Malayalam and what CDAC should be doing is to change IDN policy. We should be demanding what is our right and not blindly accpeting what is given to us. Does CDAC have a soultion to cases requiring ZWNJ? Pravs 07:20, 7 December 2010 (PST)

Conclusion

The CDAC Policy document on Malayalam IDN is not acceptable without correcting the above explained errors. In its current format, the document was prepared with lot of false assumptions and contains many technical mistakes as pointed out above. Issues introduced by careless encoding standards cannot be fixed by standarding input methods ot IDN policy. We need a consistant way of using Malayalam everywhere in digital domain. Compromise for security should not be at the cost of the language. Consultation is must with all stake holders of Malayalam Computing before preparing such an important document. This was not happened and we expect such an initiative from the authorities.

Discussion

1. Discussion about the CDACs response: http://lists.smc.org.in/pipermail/discuss-smc.org.in/2010-December/011985.html

Notes:
The latest version of this document is available at http://wiki.smc.org.in/CDAC-IDN-Critique.
1.This document is generated from http://wiki.smc.org.in/CDAC-IDN-Critique
2.Text in red color is CDAC's comments. Comments given by SMC community are given with the wiki signatures. All old comments from CDAC and SMC are retained.