Implementation of Internet Domain Names in Sinhala

Country Component	Sri Lanka	**Report no.**	
Phase no.		**Report Ref no.**	

Prepared By: **Ruwan Asanka Wasala, Chamila Liyanage, Harsha Wijayawardhana and Ruvan Weerasinghe**

Project Leader: **Dr. Ruvan Weerasinghe** **Signature:** _____

Date: **October 20, 2008**

TABLE OF CONTENTS

Implementation of Internet Domain Names in Sinhala

Asanka Wasala, Chamila Liyanage
Harsha Wijayawardhana and Ruvan Weerasinghe
University of Colombo School of Computing
35, Reid Avenue, Colombo 00700
Sri Lanka
e-mail: harsha@bit.lk,{raw,arw,cml}@ucsc.cmb.ac.lk

Abstract

This report presents the process undertaken in setting about the study and resolution of the issues surrounding the localization of Domain Names primarily for Sinhala, the methodology adopted in translating Top Level Domains (TLDs) to Sinhala, and technical issues related to the implementation of a robust localized do-main name system for Sinhala.

1. Introduction

Implementation of Internet Domain Names in Sinhala had been mooted for several years and it has become a necessity in order to popularize the use of Internet among the rural masses of Sri Lanka. Some 74% of the 20 million population of Sri Lanka count Sinhala as their first language [3]. Tamil, the second most spoken language in Sri Lanka and spoken predominantly in the North and East of the country, is the mother tongue of Tamils and a majority of Muslims in the country. Both Sinhala and Tamil languages have their own unique scripts which have been encoded in UNICODE.

In this report, the technical issues relating to the implementation of domain names in Sinhala and the methodology adopted for the translation of Top-Level Domains (TLDs) are discussed in detail. It is expected that this work will also feed into the Tamil IDN work already underway in India and Sri Lanka.

1.1 The Process

English proficiency in Sri Lanka has been estimated at only 10% prompting the Information and Communication Technology Agency (ICTA), the apex ICT policy making body in Sri Lanka, to promote the use of Sinhala and Tamil Unicode to enable the wider dissemination of ICT in the country. This effort draws from work carried out by its predecessor the Computer and Information Technology Council (CINTEC) with technical assistance from the University of Colombo. Subsequent initiatives of popularizing Internet use in local languages have been led through the active collaboration of the University of Colombo School of Computing (UCSC) and the Computer Science & Engineering Department of the University of Moratuwa (UOM). More recently, the ICTA and the LK Domain Registrar have been part of the global discussion on the fast-track Internationalized Domain Names (IDN) process. An ICTA commissioned working group, consisting of experts from within itself, the UCSC, the UOM and the LK Domain Registry (a subset of its Local Language Working Group) has been looking into issues relating to this process. The UCSC had already undertaken a study of the issues of translating generic top-level domains (gTLDs) and country-code top-level domains (ccTLDs) through the PAN Localization Project. This research is elaborated in this report. This work naturally fed into the IDN working group of the ICTA referred to above.

2. Character Set Selection

This section presents a summary of the Sinhala writing system and the allowed set of characters in a Sinhala domain name.

2.1 The Sinhala Writing System

The Sinhala character set has 18 vowels, 41 consonants and 2 semi-consonants [10] as shown in Table 1.

Vowels and corresponding vowel modifiers (within brackets): අ ආ(ා) ඇ(ැ) ඈ(ෑ) ඉ(ි) ඊ(ී) උ(ු) ඌ(ූ) ඍ(ෘ) ඎ(ෲ) ඏ(ෟ) ඐ(෠) එ(ෙ) ඒ(ේ) ඓ(ෛ) ඔ(ො) ඕ (ෝ) ඖ(ෞ) Consonants: ක ඛ ග ඝ ඞ ච ඡ ජ ඣ ඤ ඥ ඦ ට ඨ ඩ ඪ ණ ඬ ත ථ ද ධ න ඳ ප ඵ බ භ ම ඹ ය ර ල ව ශ ෂ ස හ ළ ෆ ඟ Semi-consonants: ං ඃ Special symbols: ් ා ෘ Inherent vowel remover (Hal marker): ්

Table 1: Sinhala character set

Sinhala characters are written left to right in horizontal lines. Words are delimited by a space in general. Vowels have corresponding full-character forms when they appear in an absolute initial position of a word. In other positions, they appear as 'strokes' and, are used with consonants to denote vowel modifiers. Sinhala has conjunct letters too, though they are not widely used today. A conjunct letter is formed by joining another letter to a pure consonant (e.g. ක්+ෂ=ක්ෂ). When consonants are combined with ර් - /r/ or ය් - /j/, special conjunct symbols (non-vocalic stokes) are used. ර් - /r/ immediately following a consonant can be marked by the symbol '්' - which is known as *Rakaransaya*, is added to the bottom of the consonant preceding it. Similarly, ය් - /j/, immediately following a consonant can be marked by the symbol '්‍ය' which is known as *Yansaya*, is added to the right-hand side of the consonant preceding it. The symbol '්‍ර' - /r/ known as *Repaya*, symbolizes a ර් - /r/ preceding a consonant. In addition to Repaya, Rakaransaya and Yansaya, few other consonants can be written using the conjunct form or non-conjunct form [3] [10]. These letters include ක්ෂ, ඳ, ප්ප, ඬ, ඹ, ඛ්ය, ඬ්ය, ඹ්ය, ඨ්‍ය, ඩ්‍ය, ච්‍ය and ත්‍ය.

'ෘ' - /ilu/ and 'ෲ' -/ilu:/ do not occur in contemporary Sinhala [2]. The vowels /ə/ and /ə:/, occur only in loan words of English origin. Since there are no special symbols to represent them, frequently the 'අ' vowel or 'ර' letter is used to symbolize them [Kar04].

Few letters in Sinhala can be used for writing the ancient Indic language *Pali*, the lingua franca of Buddhist Scripts. Although Pali is often written using Sinhala script, the style of writing differs from the usual writing of Sinhala. When letters are written in Pali language convention, a pure consonant (letter with Halant) followed by another consonant can be represented by writing the consonants touching each other (e.g. ක්ක) [10].

2.2 Allowed and Disallowed Characters

Sinhala character set has been standardized under Unicode and the range is U+0D80 to U+0DFF [10]. The table comprises of codes for the semi-consonants (0D82–0D83), vowels (0D85–0D96), consonants (0D9A–0DC6), the *Al-lakuna* or Halant (0DCA), vowel signs (0DCF–0DF3) and punctuation mark -*Kundaliya* (0DF4). The key decision to be made here is the allowed characters (codepoints) to be used in a Sinhala domain name.

Sinhala belongs to the Indic family of languages and the Unicode implementation of Indic languages requires the use of the Zero-Width-Joiner (ZWJ) character for encoding representations of certain words. In IDNA2003, some of the essential vowel modifiers have been disallowed in Sinhala and other Indic languages with the ZWJ being dropped in the preparation process of U-label to A-label. This was not acceptable to the majority of Indic languages including Sinhala, where the same set of codepoints with and without the ZWJ represents distinct word forms within the language. This resulted in the reformulation and release of the IDNA2008 proposal [5]by the IDN working group of ICANN. In IDNA2008, all characters in Sinhala have been allowed except for the stylistic period of Sinhala: the Kundaliya.

More importantly IDNA2008 allows the ZWJ to be accepted in context, based on rules which would be introduced into CONTEXTJ. According to the IDNA2008 specification, currently defined contextual rule only allows ZWJ to be included between two characters from the same script [5].

In Sinhala, ZWJ is used to encode conjunct letters and touching forms, albeit in a different order to that used to represent conjunct letters and ligatures. A conjunct letter is generally represented as: cons U+0DCA U+200D cons (i.e. <consonant><Sinhala Sign Al-lakuna> <ZWJ><consonant>), whereas the touching letters are represented in the code sequence: cons U+200D U+0DCA cons (i.e. <consonant><ZWJ><Sinhala Sign Al-lakuna> <consonant>). For an example, conjunct letter 'ශ්‍රී' is encoded as: U+0DB1 U+0DCA U+200D U+0DAF, and touching letters 'ක්ක' are represented as: U+0DC3 U+200D U+0DCA U+0DC3. On the basis of the above contextual analysis, the authors propose following rule[1] to be included in the IDNA2008 CONTEXTJ rules registry in order to allow both conjunct and touching letter forms in Sinhala domain names.

[1] The rule is described in the Unicode Regular Expression notation.

```
[\p(Script:Sinhala)](\u0DCA\u200D)|(\u200D\u0DCA)[\p(Script:Sinha
la)]
```

More stringent version of the above rule may test only for Sinhala consonants to be allowed on either side of `(\u0DCA\u200D)|(\u200D \u0DCA)` part. But, due to performance reasons, matching for Sinhala letters should be sufficient.

In summary, allowing all characters except the Kundaliya and the implementation of CONTEXTJ rule described above will be sufficient for successful implementation of Sinhala domain names according to the IDNA2008 protocol. In addition, the digits and hyphen should be allowed in Sinhala domain names.

3. Translating Domain Names

This section presents the translation, selection and approval methodology which was developed to unambiguously propose the most suitable terms to be used as top-level domain names (both generic and country) in Sinhala.

3.1 gTLD Translation Process

The approach of the translation process with respect to gTLDs involves initially collecting as many Sinhala terms and synonyms as possible for each English gTLD with the aid of dictionaries and thesauri [1] [4] [6] [7] [8] [9] [11] [12] [13]. In the next step, three Sinhala language scholars were consulted to ensure the correct use of the terms which were generated in the initial step. Simultaneously, appropriate short forms and abbreviations were obtained from the scholars for gTLDs, and new words were coined where no appropriate term was available in the above generation process. In the final step, each Sinhala term was carefully analyzed in terms of filtering criteria described in the section 3.1.1 and filtered above words to obtain five most appropriate Sinhala terms for each gTLD. The selected words are given in Table 2.

.com	වණිං	වාණිජ	වණික්	කොම්	විකම
.net	දැල	ජාල	ජාලය		
.info	තතු	විනැවි	පුවත්	පවත	වග
.biz	වාණිජ	වෙළඳ	සකුණ	ව්‍යාපා	ව්‍යාප
.aero	ගගන	ගුවන්	අඹර		
.asia	ආසියා	ආසි			
.cat	කැටි	කැටලන්			
.coop	සමුප	සමාගම්	සමූහ	සහය	සමා
.jobs	රැකියා	කිස	රක්සා		
.mobi	ජංගම	සර	චල	සැරි	චලන
.museum	කටුගේ	කෞ	කෞතු	කටුගෙය	පැරැණි
.name	තෙම	නාම	සිය	ස්ව	නම
.pro	ප්‍රවීණ	නිපුණ	විද්	වියති	වෘත්තික
.tel	හමුව	ටෙලි	හමු		
.travel	සංචාර	චාරිකා	සැරිසර	සැරි	ගමන
.gov	රජය				
.edu	සිප්	සතර	සරස		ඉගැනු
.mil	රණ	යුධ	හමුදා	යුද	
.int	ජගත්	ලෝ	ලොව	දියත	දැතුරු

Table 2: Selected gTLDs

3.1.1 Criteria for Selecting Best Sinhala Terms for gTLDs

Following aspects of each term have been thoroughly analyzed for proposing the best terms. In the following, a selection criterion followed by an example is given. In each of these examples, two Sinhala terms are the selected translations of the gTLD given in the brackets. Out of these two Sinhala terms, the first term satisfies the given criterion.

- Length of the term (i.e. number of letters) - absolute maximum of 6 characters: e.g. තතු vs. තොරතුරු (.info)
- Number of keystrokes - absolute maximum of 6 keystrokes: e.g. රණ vs. හමුදා (.mil)
- Typing convenience: adjacency of keys and fingers used: e.g. ගගන vs. අඹර (.aero)
- Use of Shift and AltGr keys: e.g. යුද vs. යුධ (.mil)
- Memorability: රැකියා vs. කිස (.jobs)
- Sound symbolism: සුමු vs. හවුල (.org)
- Meaningfulness: කටුගේ vs. කෞතුකාගාරය (.museum)
- Appropriateness: හමුව vs. ටෙලි (.tel)
- UNICODE storage and representation (esp. the use of U+200D - ZWJ): සිප් vs. අධ්‍යයන (.edu)
- Abbreviation possibility: සං vs. සංවිධාන (.org)
- Unambiguousness: සිප් vs. සතර (.edu)
- Frequency of usage and contemporary usage: ජාල vs. දැල (.net)
- Familiarity: කොම් vs. වණික් (.com)

3.2 ccTLD Translation Process

In selecting the most appropriate Sinhala ccTLDs, a rather more algorithmic approach was employed due to the phonetic nature of the Sinhala language. For each country name, words such as United and Island were stripped with only the most significant name retained and transliterated into Sinhala. The first two syllables of this transliterated name are extracted as the ccTLD. In cases where the transliteration above yields names consisting of two or more space separated terms, the initial syllables of each term are extracted to form the ccTLD. In either case, an absolute maximum of 6 keystrokes per ccTLD was allowed. If this process resulted in more than one possible form, an intuitive criterion was defined to propose the most suitable Sinhala ccLTD for the country concerned. This criterion takes into account the number of keystrokes, familiarity with the existing English ccLTD, and the ability to perform an intelligent guess among others. The complete list of proposed Sinhala ccLTDs is given in the Appendix.

4. Technical Issues of Implementation

4.1 Homoglyps

An analysis of a 10 million word corpus of Sinhala shows that characters: U+0D8E - ඎ /rii/, U+0D8F - ඏ /ilu/, U+0D90 - ඐ /iluu/, U+0DA6 - ඦ /^{n}ʤə/ have occurred zero times. This result shows that the use of above characters in Sinhala domain names may not be necessary. Especially, the character U+0D8E - ඎ looks similar to U+0DB4 - ප /pə/ which is a frequently used character.

A careful visual inspection of Sinhala letters revealed that some of the characters of Sinhala Alphabet are prone to phishing attacks due to their visual similarities, especially when they are displayed in the default address bar of browsers. These characters are presented in the Table 6.

In addition, shapes of several Sinhala characters are found to be visually similar to letters in other Indic languages. These characters are presented in the Table 7.

Solutions to homoglyph attacks may range from simple n-gram or dictionary based algorithms to advance natural language processing and image processing techniques. However, such fully-automated approaches are impractical, or ineffective in the domain name registration process. Therefore, policies should be formulated in order to reduce the vulnerability of homoglyph attacks. IDNA is an end-application focused solution. Especially, browsers play a major role in IDN implementation. Two simple solutions to reduce the risk of phishing include prohibiting the use of stylistic fonts when displaying IDNs and increasing the font size of address bar of browsers. A good feature to implement in browsers is that a pop-up window to display the IDN in a bigger font size upon the identification of the IDN (by ACE prefix or Unicode characters) as it is being entered into address bar of a browser. Therefore, providing a feature such as the above will help user to distinguish and recognize characters easily. The other possibility is that, local registrars can implement a policy which allows users to reserve or block IDNs with obvious homoglyphs, for a small marginal pay. This solution will help to overcome any processing overheads in the registrars' end.

Visually Identical Characters	Visually Identical Term(s)	Correct Term(s)	Visually Identical Term(s)	Correct Term(s)
ඛ - බ	ඛබාට-නමක්	බබාට-නමක්	ඛබාට-නමක්	බබාට-නමක්
ඩ - ඩ	පාඩම්	පාඩම්	පාඬම්	පාඩම්
ඡ - ජ	අංඡනම	අංජනම	අංෙනම	අංජනම
ප - ප	කැඩපත	කැඩපත	කැඬපත	කැඩපත
සෘ - සෘ	සෘණ	සෘණ	සෘණ	සෘණ
ඹ - ඹ	කොළඹ	කොළඹ	කොළඹ	කොළඹ
හ - හ	අහස	අහස	අහස	අහස
හ - හ	ඉහත	ඉහත	ඉඟත	ඉහත
ච - ව	සචන	සවන	සචන	සවන
ඡ - ජ	සමාඡය	සමාජය	සමාඡය	සමාජය
එ - එ	එතෙර	එතෙර	එතෙර	චතෙර
ළු - එ	ළුළකිරි	එළකිරි	ළුළකිරි	චිළකිරි
එ - ළු	කළුවර	කළුවර	කළුවර	කළුවර
ඬ - ඬ	කණ්ඩලම	කණ්ඩලම	කණ්ඩලම	කණ්ඩලම
ඬ - ඬ	අඬහැරය	අඬහැරය	අඬහැරය	අඬහැරය
ඬ - ඬ	කඬවර	කඬවර	කඬවර	කඬවර
ම් - ම	අම්රදේව	අමරදේව	අම්රදේව	අමරදේව
ධ් - ධ	සද්ධර්මය	සද්ධර්මය	සද්ධ්ර්මය	සද්ධර්මය
බ් - බ	ප්‍රබල	ප්‍රබල	ප්‍රබ්ල	ප්‍රබල
ව් - ව	නවතම	නවතම	නව්තම	නවතම
ට් - ට	වැටිකොලු	වැටකොලු	වැට්කොලු	වැටකොලු
ව් - ව	වනව්ර	වනවර	වනව්ර	වනවර
ක්‍ඩ - ක්‍ඩ	නිරුජ්ක්‍ඩති	නිරුජ්ක්‍ඩති	නිරුජ්ක්‍ඩ්ති	නිරුජ්ක්‍ඩති
ඔ - ඕ	ඔලු	ඕලු	ඔලු	ඕලු
ඩ් - ඩ	ලඬකාව	ලඬකාව	ලඬ්කාව	ලඬකාව
ඩ් - ඩ්	පොඩ්ඩක්	පොඩ්ඩක්	පොඬ්ඩක්	පොඩ්ඩක්
ජී - ජී	අව්චිඡීන්න	අව්චිඡීන්න	අව්චිඡින්න	අව්චිඡින්න
ණී - ණී	පැරණී	පැරණී	පැරණී	පැරණී
ටී - ටී	අස්ටී	අස්ටී	අස්ටී	අස්ටී
ඵී - ඵී	සුපිප්ඵීත	සුපිප්ඵීත	සුපිප්වීත	සුපිප්වීත
ටූ - ටූ	ඇහැටූ	ඇහැටූ	ඇහැට්ූ	ඇහැටූ
ණූ - ණූ	පණූවා	පණූවා	පණූවා	පණූවා
ඤූ - ඤූ	සබ්බඤ්ඤූ	සබ්බඤ්ඤූ	සබ්බඤ්ඤූ	සබ්බඤ්ඤූ
කූ - කූ	කූරුබිලිය	කූරුබිලිය	කූරුබිලිය	කූරුබිලිය
පූ - පූ	පූනර්භවය	පූනර්භවය	පූනර්භවය	පූනර්භවය
ී - ී	ධරණීතලය	ධරණීතලය	ධරණීතලය	ධරණීතලය
ෟ - ෟ				

Table 6: Homoglyphs Within the Sinhala Script

Script	Glyph	Confusing Sinhala Glyph
Bengali	ঌ (U+098C)	ඓ (U+0DC0)
Gujarati	ૐ (U+0AB3)	ගූ (U+0D9C) (U+0DD4)
Gujarati	ઋ (U+0A9D)	ඳ් (U+0DB1) (U+0DCA)
Telugu	ఎ (U+0C0E)	ඛ (U+0DB6)
Telugu	ఘ (U+0C28)	ඛ් (U+0DB6) (U+0DCA)
Telugu	ఒ (U+0C2C)	ඝ (U+0DC3)
Telugu	ర (U+0C30)	ර් (U+0DBB) (U+0DCA)
Telugu	ళ (U+0C35)	ඛ් (U+0DB6) (U+0DCA)
Kannada	ಎ (U+0C8E)	ඨ (U+0DB0)
Kannada	ಒ (U+0CAC)	ඝ (U+0DC3)
Malayalam	ക (U+0D15)	ක (U+0D9A)
Malayalam	ഗ (U+0D17)	ග (U+0D9C)
Malayalam	ന (U+0D28)	ඉ (U+0DC6)

Table 7: Homoglpys in Other Scripts

4.2 Mixing of Scripts

Though IDN2008 fast-track process prohibits mixing scripts in an internationalized domain names, there is a possibility that future revisions may allow mixing of scripts.

Since Sinhala and Tamil are both national languages in Sri Lanka, authors had looked deeply into the mixing of both languages in the registration of Internet Domains. It was found that Tamil and Sinhala will not have multi letter homoglyphs. Thus, mixing of Tamil and Sinhala characters should be allowed, as it would not cause any problems. However, mixing Latin script with Sinhala script pose a phishing threat as English letter 'o' may looks similar to Sinhala letter Anusvaraya 'ං'. (e.g. as in http://ශ්‍රීලංකන්.ගුවන්, where third letter is actually simple letter 'o'). There exist two solutions to this problem. First one is to define a CONTEXTJ rule to restrict the use of English letter 'o' in between Sinhala characters. The second one is to allow user to reserve or block the mixed script domain name which has letter 'o' in between Sinhala characters, by paying a little extra cost. However, owing to technical implementation difficulties associated with the first option, the authors recommend the latter option.

4.3 Multiple Forms of the Same Word (Phonetic Variants)

In Sinhala, certain words can be written in many forms like in other Indic and Arabic languages. These words can be further divided into four categories based on their method of formation.

4.3.1 Existence of Conjunct Form and Non-Conjunction Form

Several words can be written in their conjunct forms and non-conjunct forms. Some examples are given below: (Conjunct letter is included in the first word) {ක්‍ෂේත්‍රය, ක්ෂේත්‍රය}, {වඳනාව, වන්දනාව}, {සඤ්ඣ, සන්ධ්‍}.

4.3.2 Words with Yansaya, Rakaransaya and Repaya (Reph)

A comprehensive analysis of a ten million word corpus revealed that certain words with Yansaya, Rakaransaya and Repaya[2] may appear in different orthographic representations. These characters are encoded with the use of the ZWJ.

The Rakaransaya is encoded by the codepoint sequence: <consonant> U+0DCA U+200D U+0DBB. Similarly, The Yansaya is represented by the codepoint sequence <consonant> U+0DCA U+200D U+0DBA. The Repaya is encoded as U+0DBB U+0DCA U+200D <consonant>, where U+200D is the ZWJ [10]. In certain situations one can elect not to use the above non-vocalic strokes but to use their underlying representation of letters 'ර' and 'ය'. In this case, the ZWJ should be omitted. (e.g. රත්‍රන් and රත්‍රන්, both form are valid and acceptable, similarly both විද්‍යුත් and විද්‍යුත් are considered valid and acceptable). As such, a word comprised of above character can

[2] See Section 2.1 for the definitions of Yansaya, Rakaransaya and Repaya.

be written in two different forms if desired, though the general practice is to use the conjunct symbols: Rakaransaya and Yansaya.

The Repaya is used to represent the letter 'ර' preceding a consonant. Use of the Repaya is optional and it is not used in modern script, though one may decide to use Repaya in a domain name (e.g. both ධර්‍මය and ධර්මය are valid, where ධර්‍මය is written using the Repaya). Furthermore, the pair of letters 'ර්ය' can be written in three different ways when second letter 'ය' is not associated with vowel modifiers: 'ේ' /e:/ and 'ෝ' /o:/ (i.e. 'යේ', 'යෝ'). The three forms are 'ය්‍ය', 'ය්' and 'ර්ය' where last form is the popular contemporary usage form (e.g. ආය්‍ය, ආය්, ආර්ය all are valid and represent the same word and the meaning).

4.3.3 Touching form for PALI

One may want to retain either the classical Pali touching letter form exactly as found in sacred Buddhist books (e.g. බුඬ) or in separate letter form (e.g. බුද්ධ) in a Sinhala domain name.

It is worthy to mention that some pairs of letters can be written in all three forms: normal, touching and conjunct (e.g. ද්ධ , ඬ and ඩ).

4.3.4 Spelling Variants

Some spelling variants are considered equally correct. Phonetic similarity provides clues to detect real-word spelling variants. Few examples for this category include, {ලඬකාව, ලංකාව}, {කෂාය, කසාය}, {කාංසාව, කාංශාව}, {පංචය, පඤ්චය} and {ඥෙයා, ඤෙයා}.

The ability to write same words in different forms cause serious IDN implementation problems. For an example, the word ධර්‍මාචාර්ය can take up to 6 valid forms: {ධර්‍මාචාය්‍ය, ධර්‍මාචාර්ය, ධර්මාචාය්‍ය, ධර්මාචාර්ය, ධර්‍මාචාය්, ධර්මාචාය්}. When combined several such words together as in a domain name, may produce unpredictable number of valid domain names as explained by the following example:

- Valid forms of the 1st word: { හර්ෂගේ, හර්‍ෂගේ} meaning Harsha's *(Harsha is a person name)*
- Valid forms of the 2nd word: {කාර්‍යාලය, කාය්‍යාලය, කාර්යාලය} meaning office
- Possible valid combined forms:

  {හර්ෂගේකාර්‍යාලය, හර්ෂගේකාය්‍යාලය, හර්ෂගේකාර්යාලය, හර්‍ෂගේකාර්‍යාලය, හර්‍ෂගේකාය්‍යාලය, හර්‍ෂගේකාර්යාලය} meaning Harsha's office.

Every domain name has a business value, and thus it is necessary to establish a policy with regard to words with multiple forms. The authors propose following policy to be implemented.

The policy should allow user to request the preferred form among the multiple forms, from the domain name registrar. Then, the domain registrar can register the requested domain name if it satisfies other policies and conditions. Upon proving the requested domain name has multiple forms, the registrar should allow the user to reserve each additional form by paying a reasonable additional cost. This policy will reduce the burden of the domain name registrar by forcing the user to deal with the words that have multiple forms.

4.4 Issues in Browsers

A test was designed to determine the level of various browsers' support for Internationalized Domain Names. Especially to identify the issues related to manipulation of Sinhala domain names in commonly used browsers.

The test consists of entering two (non-existent) Sinhala domain names (U-labels), and two valid A-labels into the default address bar of several popular browsers, and examine the behavior of browsers in manipulating Sinhala domain names. Test cases are given in the Table 3.

Test No.	Sinhala Domain Name	Description
1	http://සරසවි.lk	U-label that does not contain ZWJ.
2	http://ශ්‍රීලංකා.gov	U-label that contains ZWJ.
3	http://xn--10ckhb0f.lk	A-label generated from a string that does not contain ZWJ (i.e. A-Label equivalent to http://සරසවි.lk)
4	http:// xn--fzc2c3eis8cxb2a6013b.gov	A-label generated from a string that contains ZWJ. (i.e. A-Label equivalent to http:// http://ශ්‍රීලංකා.gov)

Table 3. Sinhala Domain Name Test Cases for Browsers

The above tests were designed to identify whether CONTEXTJ rules have been implemented in browsers or not. In addition, the tests were also useful in identifying rendering issues of Sinhala characters in browsers.

Test results have been divided into two parts and given in Table 4 and Table 5. Table 4 summarizes the results of Test #1 and Test #2, while Table 5 summarizes the results of Test #3 and Test #4.

Browser		Typing/Rendering Issues in Address Bar	A-label	Label displayed in the address bar
Firefox version 3.0.1	Test 1	None	http://xn--10ckhb0f.lk	http://xn--10ckhb0f.lk
	Test 2	None	http://xn--fzc2c3eis8cxb2a.gov	http://xn--fzc2c3eis8cxb2a.gov
Internet Explorer 8 (Beta)	Test 1	None	http://xn--10ckhb0f.lk	http://xn--10ckhb0f.lk
	Test 2	None	http://xn--fzc2c3eis8cxb2a.gov	http://xn--fzc2c3eis8cxb2a.gov
Safari 1.3.1	Test 1	None	http://xn--10ckhb0f.lk/	http://xn--10ckhb0f.lk
	Test 2	Does not render characters with ZWJ. e.g.ශ්‍රීලංකා.gov	http://xn--fzc2c3eis8cxb2a.gov	http://xn--fzc2c3eis8cxb2a.gov
Opera 9.6 (Beta)	Test 1	None	http://xn--10ckhb0f.lk	http://xn--10ckhb0f.lk
	Test 2	Does not render characters with ZWJ. e.g. ශ්‍රීලංකා.gov	http://xn--fzc2c3eis8cxb2a.gov	http://xn--fzc2c3eis8cxb2a.gov
Google Chrome 0.2.149.29	Test 1	Extra spaces in domain name. e.g. ස ර ස වි. lk	http://xn--10ckhb0f.lk	http://xn--10ckhb0f.lk
	Test 2	Extra spaces in domain name e.g. ශ්‍රීලංකා. gov	http://xn--fzc2c3eis8cxb2a.gov	http://xn--fzc2c3eis8cxb2a.gov

Table 4: Test #1 and Test #2 : U-label to A-Label Conversion Process

Browser		Label displayed in the address bar/Browser response
Firefox version 3.0.1	Test 3	http://xn--10ckhb0f.lk
	Test 4	http://xn--fzc2c3eis8cxb2a6013b.gov
Internet Explorer 8 (Beta)	Test 3	http://xn--10ckhb0f.lk/
	Test 4	Gives message: http://xn--fzc2c3eis8cxb2a6013b.gov/ is currently unavailable
Safari 1.3.1	Test 3	http://xn--10ckhb0f.lk/
	Test 4	http://xn--fzc2c3eis8cxb2a6013b.gov
Opera 9.6 (Beta)	Test 3	http://xn--10ckhb0f.lk
	Test 4	Gives message: The URL http://xn--fzc2c3eis8cxb2a6013b.gov/ contains characters that are not valid in the location they are found.
Google Chrome 0.2.149.29	Test 3	http://xn--10ckhb0f.lk
	Test 4	http://xn--fzc2c3eis8cxb2a6013b.gov

Table 5: A-label to U-label Conversion Process

Common observation in all browsers is that, a browser will always convert a U-label into an A-label and display the corresponding A-label in the address bar. When an A-label is entered in the address bar (as in the test cases #3 and #4), ideally a browser should convert it into corresponding U-label by default [Kle08]. However, this did not happen in the browsers chosen for this test. A simple *hack*[3] will enable A-label to U-label conversion process in Firefox browser. Firefox maintains a 'white-list' of IDNA supportive scripts. Once Sinhala is included into that list, Firefox converts Sinhala A-labels into U-labels and displays U-labels in the address bar. However, even with this modification, Firefox failed the Test #4 - as it contained a ZWJ character. The above tests revealed that current browsers considered in these tests cannot manipulate IDNs with ZWJ. Thus, from these tests, it can be concluded that still all most all popular browsers do not confirm to IDNA2008 implementation, hence they lack support for Sinhala domain names.

IDNA is an application driven process. Therefore, in order to implement internationalized domain names successfully, these applications must conform to the latest revisions of IDNA2008 protocol. Furthermore, any script specific issues such as rendering problems have to be solved. Browsers play a major role in implementing IDNA. Therefore, web browser developers should frequently modify their browsers to accommodate the changes introduced in the latest revisions of IDNA protocol.

[3] See http://www.mozilla.org/projects/security/tld-idn-policy-list.html for details.

5. Conclusion and Future Work

It has been planned to implement International domain names in Sinhala by the early 2009. Although it is technically possible to implement Sinhala domain names by that time, many policies have to be formulated and certain technical issues have to be ironed out for the successful implementation of Sinhala domain names as pointed out in the report. Although some work had been carried out for the identification of technical and non-technical issues for the implementation of IDNs, certain issues have yet to be delved deeply such as translation issues for secondary level domain names. Therefore future work is necessary for the following:

- Translation issues - Secondary level domain names
- Translation of resources of Internet such as www, ftp etc from English to Sinhala
- Policies on rude words and the use of slang words in IDN
- And to formulate continuous monitoring process for issues relating to IDNs and new domain names.

In concluding, this report describes in detail a robust process identified and undertaken to arrive at a stable and acceptable solution to Internationalized Domain Names in Sinhala. It also enumerates the main issues to be faced and resolved in the IDN process for scripts of non-Latin languages. In particular it sets out a strategy and a systematic process, considering both human and technical factors, for providing a sustainable IDN solution for languages especially belonging to the Indic family.

Acknowledgement

The authors would like to thank Sinhala Language scholars Prof. Thissa Jayawardane, Aelian de Silva and Prof. Sucharitha Gamlath for their invaluable support and advice throughout the study. We also wish to acknowledge the contribution of Mr. Namal Udalamaththa, Mr. Randil Pushpananda and Mr. Asiri Ranasinghe of Language Technology Research Laboratory of the University of Colombo School of Computing, Sri Lanka.

References

1. Sandagomi Coperahewa, *Dictionary of Sinhala Spelling*, S. Godage and Brothers, 2000, ISBN 955-20-4462-6.
2. J. B. Disanayaka, *The Structure of Spoken Sinhala*, National Institute of Education, 1991, ISBN 955-579-053-X.
3. W. S. Karunatillake, *An Introduction to Spoken Sinhala*, M. D. Gunasena & Company Limited, 2004, ISBN 955-21-0878-0.
4. W. S. Karunatillake, *Pada Manjari (Samanartha Pada Sangrahaya)*, S. Godage and Brothers, 2007, ISBN 955-20-9946-6.
5. J. Klensin, *Internationalized Domain Names for Applications (IDNA): Definitions, Background and Rationale*, Internet-Draft, Work – In - Progress, Web site, http://www.ietf.org/ internet-drafts/draft-ietf-idnabis-rationale-03.txt, October 7, 2008.
6. G. P. Malalasekera, *English – Sinhala Dictionary*, M. D. Gunasena & Company Limited, 2007, ISBN 978-955-21-1674-2.
7. Aelian De Silva, *Technical Terms in Sinhala*, The Modern Book Company, 2002, ISBN 955-558-083-9.
8. W. Sorata, *Sri Sumangala Sabdakosaya*, S. Godage and Brothers, 1999, Part 2.
9. W. Sorata, *Sri Sumangala Sabdakosaya*, S. Godage and Brothers, 2006, Part 1.
10. Sri Lanka Standard, *SLS 1134:2004 - Sinhala Character Code for Information Interchange,* Sri Lanka Standards Institute, 2004, Revision 2.
11. Harischandra Wijayatunga, *Practical Sinhala Dictionary*, Ministry of Culturel Affairs, 1982, Vol 1.
12. Harischandra Wijayatunga, *Practical Sinhala Dictionary*, Ministry of Culturel Affairs, 1984, Vol 2.
13. Harischandra Wijayatunga, *Gunasena Great Sinhala Dictionary*, M. D. Gunasena & Company Limited, 2005, ISBN 955-21-1423-3.

Appendix

Proposed Sinhala Country-Code Top-Level Domain Names

ccTLD	Description	Sinhala Term
.ac	Ascension Island	ඇසෙ
.ad	Andorra	ඇඩෝ
.ae	United Arab Emirates	අරාබි
.af	Afghanistan	ඇෆ්ග
.ag	Antigua and Barbuda	ඇබා
.ai	Anguilla	ඇන්ගු
.al	Albania	ඇල්
.am	Armenia	ආමේ
.an	Netherlands Antilles	ඇන්
.ao	Angola	ඇගෝ
.aq	Antarctica	ඇටා
.ar	Argentina	ආජ
.as	American Samoa	ඇමස
.at	Austria	ඔට්ර
.au	Australia	ඔස්
.aw	Aruba	අරූබා
.ax	Aland Islands	ඇලන්
.az	Azerbaijan	අසර්
.ba	Bosnia and Herzegovina	බොහ
.bb	Barbados	බාබ
.bd	Bangladesh	බංගලි
.be	Belgium	බෙජි
.bf	Burkina Faso	බර්කි
.bg	Bulgaria	බගේ
.bh	Bahrain	බහරේ
.bi	Burundi	බුරු
.bj	Benin	බෙනී
.bm	Bermuda	බම්
.bn	Brunei Darussalam	බෘද
.bo	Bolivia	බොලි
.br	Brazil	බ්‍රසී
.bs	Bahamas	බහමා
.bt	Bhutan	භූතා
.bv	Bouvet Island	බෝවෙ
.bw	Botswana	බොවා
.by	Belarus	බෙල
.bz	Belize	බෙලී
.ca	Canada	කැනඩා
.cc	Cocos (Keeling) Islands	කොකො
.cd	Congo, The Democratic Republic of the	කොප්‍ර

ccTLD	Description	Sinhala Term
.cf	Central African Republic	මධ්‍ර
.cg	Congo, Republic of	කොගො
.ch	Switzerland	ස්විස්
.ci	Cote d'Ivoire	කෝඩ්
.ck	Cook Islands	කුක්
.cl	Chile	චිලී
.cm	Cameroon	කැම
.cn	China	චීන
.co	Colombia	කොලෝ
.cr	Costa Rica	කොරී
.cu	Cuba	කියුබා
.cv	Cape Verde	කේව
.cx	Christmas Island	ක්‍රිස්
.cy	Cyprus	සයිප්‍ර
.cz	Czech Republic	චෙක්
.de	Germany	ජම
.dj	Djibouti	ජිබූ
.dk	Denmark	ඩෙන්මා
.dm	Dominica	ඩොම්
.do	Dominican Republic	ඩොමිජ
.dz	Algeria	ඇජී
.ec	Ecuador	ඉක්ව
.ee	Estonia	එටෝ
.eg	Egypt	ඊජි
.eh	Western Sahara	බසහ
.er	Eritrea	එරි
.es	Spain	ස්පා
.et	Ethiopia	ඉති
.eu	European Union	යුරෝ
.fi	Finland	පින්ල
.fj	Fiji	ෆිජී
.fk	Falkland Islands (Malvinas)	ෆෝල
.fm	Micronesia, Federated States of	මයි
.fo	Faroe Islands	ෆෙරෝ
.fr	France	ප්‍රංශ
.ga	Gabon	ගැබොා
.gb	United Kingdom	එරා
.gd	Grenada	ග්‍රැනඩා
.ge	Georgia	ජෝජී
.gf	French Guiana	ෆ්‍රෙග
.gg	Guernsey	ගුන්සි
.gh	Ghana	ඝානා
.gi	Gibraltar	ජිබ්‍රෝ
.gl	Greenland	ග්‍රීන්
.gm	Gambia	ගැම්බි
.gn	Guinea	ගිනියා

ccTLD	Description	Sinhala Term
.gp	Guadeloupe	ග්වාඩ්
.gq	Equatorial Guinea	ඉගිනි
.gr	Greece	ග්‍රීසි
.gs	South Georgia and the South Sandwich Islands	දඡෝ
.gt	Guatemala	ගෝත
.gu	Guam	ගුවාම්
.gw	Guinea-Bissau	ගිනිබි
.gy	Guyana	ගයනා
.hk	Hong Kong	හොං
.hm	Heard and McDonald Islands	හමැ
.hn	Honduras	හොඩූ
.hr	Croatia/Hrvatska	ක්‍රොඒ
.ht	Haiti	හයිටි
.hu	Hungary	හංගේ
.id	Indonesia	ඉන්දූ
.ie	Ireland	අයර්
.il	Israel	ඊශ්‍රා
.im	Isle of Man	අමැ
.in	India	ඉන්දි
.io	British Indian Ocean Territory	බ්‍රිසාක
.iq	Iraq	ඉරාක
.ir	Iran, Islamic Republic of	ඉරාන
.is	Iceland	අයිල
.it	Italy	ඉතාලි
.je	Jersey	ජර්සි
.jm	Jamaica	ජැමෙයි
.jo	Jordan	ජෝදා
.jp	Japan	ජපාන
.ke	Kenya	කෙන්යා
.kg	Kyrgyzstan	කිරිගි
.kh	Cambodia	කාම්
.ki	Kiribati	කිරිබැ
.km	Comoros	කොමෝ
.kn	Saint Kitts and Nevis	කිනේ
.kp	Korea, Democratic People's Republic	කොප්‍ර
.kr	Korea, Republic of	කොරි
.kw	Kuwait	කුවේට
.ky	Cayman Islands	කේම
.kz	Kazakhstan	කසක
.la	Lao People's Democratic Republic	ලාඕ
.lb	Lebanon	ලෙබන
.lc	Saint Lucia	ශාලූ
.li	Liechtenstein	ලයිවෙ
.lk	Sri Lanka	ලක / ලංකා

ccTLD	Description	Sinhala Term
.lr	Liberia	ලයිබ්
.ls	Lesotho	ලෙසතො
.lt	Lithuania	ලිතු
.lu	Luxembourg	ලක්සම්
.lv	Latvia	ලැට්ව්
.ly	Libyan Arab Jamahiriya	ලිබිජම
.ma	Morocco	මොරො
.mc	Monaco	මොනා
.md	Moldova, Republic of	මොල්
.me	Montenegro	මොට්
.mg	Madagascar	මැඩග
.mh	Marshall Islands	මාෂ
.mk	Macedonia, The Former Yugoslav Republic of	මැසි
.ml	Mali	මාලි
.mm	Myanmar	මියන්
.mn	Mongolia	මොන්
.mo	Macao	මැකා
.mp	Northern Mariana Islands	නෝම
.mq	Martinique	මාටිනි
.mr	Mauritania	මොරි
.ms	Montserrat	මොන්ස
.mt	Malta	මෝටා
.mu	Mauritius	මොරි
.mv	Maldives	මාලදි
.mw	Malawi	මලාව්
.mx	Mexico	මෙක්සි
.my	Malaysia	මැලේ
.mz	Mozambique	මොසැ
.na	Namibia	නැම්බ්
.nc	New Caledonia	නිකැල
.ne	Niger	නයිජ
.nf	Norfolk Island	නෝෆ
.ng	Nigeria	නයිජී
.ni	Nicaragua	නිකරැ
.nl	Netherlands	නෙදර්
.no	Norway	නෝවෙ
.np	Nepal	නේපාල
.nr	Nauru	නවුරු
.nu	Niue	නිවූ
.nz	New Zealand	නවසී
.om	Oman	ඕමාන
.pa	Panama	පැනමා
.pe	Peru	පේරු
.pf	French Polynesia	ප්‍රපො
.pg	Papua New Guinea	පැනිව්
.ph	Philippines	පිලිපි

ccTLD	Description	Sinhala Term
.pk	Pakistan	පකිතාන
.pl	Poland	පෝල
.pm	Saint Pierre and Miquelon	සාපිම
.pn	Pitcairn Island	පිට්ක
.pr	Puerto Rico	පුරි
.ps	Palestinian Territory, Occupied	පලස්
.pt	Portugal	පෘතු
.pw	Palau	පලාවු
.py	Paraguay	පැරගු
.qa	Qatar	කටාර්
.re	Reunion Island	රියුනි
.ro	Romania	රොමේ
.rs	Serbia	සර්බි
.ru	Russian Federation	රුසියා
.rw	Rwanda	රුවඩා
.sa	Saudi Arabia	සෞදි
.sb	Solomon Islands	සොල
.sc	Seychelles	සීෂෙ
.sd	Sudan	සුඩාන
.se	Sweden	ස්වීඩ
.sg	Singapore	සිංග
.sh	Saint Helena	ශාහෙ
.si	Slovenia	ස්ලෝ
.sj	Svalbard and Jan Mayen Islands	ස්වජා
.sk	Slovak Republic	ස්ලොව
.sl	Sierra Leone	සිලි
.sm	San Marino	සැන්ම
.sn	Senegal	සෙනගාල
.so	Somalia	සෝලි
.sr	Suriname	සූරිනා
.st	Sao Tome and Principe	සැටෝ
.su	Soviet Union (being phased out)	සෝස
.sv	El Salvador	එල්සැ
.sy	Syrian Arab Republic	සිරි
.sz	Swaziland	ස්වාසි
.tc	Turks and Caicos Islands	ටක
.td	Chad	චැඩ්
.tf	French Southern Territories	ප්‍රදේ
.tg	Togo	ටෝග
.th	Thailand	තායි
.tj	Tajikistan	ටැජි
.tk	Tokelau	ටොකෙ
.tl	Timor-Leste	ටිලෙ
.tm	Turkmenistan	ටක්මේ
.tn	Tunisia	ටියු
.to	Tonga	ටොන්

ccTLD	Description	Sinhala Term
.tp	East Timor	නැටි
.tr	Turkey	තුර්කි
.tt	Trinidad and Tobago	ට්‍රිටෝ
.tv	Tuvalu	ටුවාලු
.tw	Taiwan	තායිවා
.tz	Tanzania	ටැන්සා
.ua	Ukraine	යුක්‍රේ
.ug	Uganda	උගන්ඩා
.uk	United Kingdom	එංග
.um	United States Minor Outlying Islands	එජදූ
.us	United States	එජ
.uy	Uruguay	උරුගු
.uz	Uzbekistan	උස්බෙ
.va	Holy See (Vatican City State)	වතිකා
.vc	Saint Vincent and the Grenadines	විග්‍රෙන
.ve	Venezuela	වෙනිසි
.vg	Virgin Islands, British	වබ්‍රි
.vi	Virgin Islands, U.S.	වඇමරි
.vn	Vietnam	වියට්
.vu	Vanuatu	වනුවා
.wf	Wallis and Futuna Islands	වොෆ
.ws	Samoa	සැමෝ
.ye	Yemen	යේම
.yt	Mayotte	මයෝ
.yu	Yugoslavia	යුගො
.za	South Africa	දඉ
.zm	Zambia	සැම්බි
.zw	Zimbabwe	සිබා