

# Implementation of Internet Domain Names in Sinhala

Harsha Wijayawardhana, Asanka Wasala,  
Ruvan Weerasinghe and Chamila Liyanage  
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 paper 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 domain name system for Sinhala.

**Keywords:** Sinhala Domain Names, Internationalized Domain Names, IDNA, Sinhala gTLDs

## 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 [Kar04]. 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 paper, 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 – an IDRC funded project bringing together 10 Asian countries in the region looking into various aspects of Localization. This work naturally fed into the IDN working group of the ICTA referred to above.

The rest of this paper is organized as follows: Section 2 gives an overview of the Sinhala writing system and the allowed characters to be a



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 [Kle08] 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 [Kle08].

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 ‘`ඤ`’ is encoded as: `U+0DB1 U+0DCA U+200D U+0DAF`, and touching letters ‘`ඤඤ`’ are represented as: `U+0DC3 U+200D U+0DCA U+0DC3`. On the basis of the above contextual analysis, the authors propose following rule<sup>1</sup> to be included in the IDNA2008 CONTEXTJ rules registry in order to allow both conjunct and touching letter forms in Sinhala domain names.

```
[\p{Script:Sinhala}](\u0DCA\u200D)|
(\u200D\u0DCA)[\p{Script:Sinhala}]
```

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

---

<sup>1</sup> The rule is described in the Unicode Regular Expression notation.

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 [Cop00] [Kar07] [Mal07] [Sil02] [Sor99] [Sor06] [Wij82] [Wij84] [Wij05]. 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 of the Appendix.

##### 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. නතු 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. ගනන 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. වෙලි (.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: කොමි 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 con-

cerned. 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 available at: <http://www.ucsc.cmb.ac.lk/ltr1/pub1ic/cc1td.htm>.

## 4 Technical issues of implementation

### 4.1 Homoglyphs

An analysis of a 10 million word corpus of Sinhala shows that characters: U+0D8E - ඩෘ /rii/, U+0D8F - ඩ /ilu/, U+0D90 - ඩෘ /iluu/, U+0DA6 - ජ /<sup>n</sup>ඳෘ/ 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 - ඩ looks similar to U+0DB4 - ඩ /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 of the Appendix. 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 of the Appendix.

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.

## 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 'ඔ'. (e.g. as in [http:// ශ්‍රීලංකන්.ඉවන්](http://ශ්‍රීලංකන්.ඉවන්), 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-conjunct 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) {කේශ්‍රය, ක්ෂේශ්‍රය}, {වඤ්චාව, වන්ඤ්චාව}, {සකී, සන්ධි}.

### 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<sup>2</sup> 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 [Sta04]. In certain situations one can elect not to use the above non-vocalic strokes but to use their underlying representation of letters 'ඊ' and 'ඔ'. In this case, the ZWJ should be omitted. (e.g. ඊඊඊ and ඊඊඊ, both form are valid and acceptable, similarly both චිඳ්‍රන් and චිඳ්‍රන් are considered valid and acceptable). As such, a word comprised of above character can 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 'ඊ' 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 ධම්ම and ධර්මම are valid, where ධම්ම is written using the Repaya). Furthermore, the pair of letters 'ඊඔ' can be written

<sup>2</sup> See Section 2.1 for the definitions of Yansaya, Rakaransaya and Repaya.

in three different ways when second letter ‘ය’ is not associated with vowel modifiers: ‘ඒ’ /e:/ and ‘ඔ’ /o:/ (i.e. ‘යෙ’, ‘යෝ’). The three forms are ‘ය්’, ‘යී’ and ‘යූ’ where last form is the popular contemporary usage form (e.g. ආය්, ආයී, ආයූ 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. බුද්ධ) 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, ආ and ධ).

### 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, {ලඛකාව, ලංකාව}, {කෂාය, කසාය}, {කාංඝාව, කාංඝාව}, {පංචය, පඤ්චය} and {ඤ්ජය, ඤ්ජය}.

The ability to write same words in different forms cause serious IDN implementation problems. For an example, the word ධර්මාවාච්ඡා can take up to 6 valid forms: {ධර්මාවාච්ඡා, ධර්මාවාච්ඡා, ධර්මාවාච්ඡා, ධර්මාවාච්ඡා, ධර්මාවාච්ඡා, ධර්මාවාච්ඡා}. 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, කාර්යාලය, කාර්යාලය } meaning office
- Possible valid combined forms:
  - { හර්ෂගේකාර්යාලය, හර්ෂගේකාර්යාලය, හර්ෂගේකාර්යාලය, හර්ෂගේකාර්යාලය, හර්ෂගේකාර්යාලය, 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 of the Appendix.

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.

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]. How-

ever, this did not happen in the browsers chosen for this test. A simple *hack*<sup>3</sup> 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.

## 5 Conclusion and future directions

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 paper. 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 paper 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

This work has been supported through the PAN Localization Project ([www.PANL10n.net](http://www.PANL10n.net)), grant from the International Development Research Center (IDRC), Ottawa, Canada, administered through the Center for Research in Urdu Language Processing, National University of Computer and Emerging Sciences, Pakistan. 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

- [Cop00] Sandagomi Coperahewa, *Dictionary of Sinhala Spelling*, S. Godage and Brothers, 2000, ISBN 955-20-4462-6.

---

<sup>3</sup> See <http://www.mozilla.org/projects/security/tld-idn-policy-list.html> for details.

- [Dis91] J. B. Disanayaka, *The Structure of Spoken Sinhala*, National Institute of Education, 1991, ISBN 955-579-053-X.
- [Kar04] W. S. Karunatillake, *An Introduction to Spoken Sinhala*, M. D. Gunasena & Company Limited, 2004, ISBN 955-21-0878-0.
- [Kar07] W. S. Karunatillake, *Pada Manjari (Samanartha Pada Sangrahaya)*, S. Godage and Brothers, 2007, ISBN 955-20-9946-6.
- [Kle08] 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.
- [Mal07] G. P. Malalasekera, *English - Sinhala Dictionary*, M. D. Gunasena & Company Limited, 2007, ISBN 978-955-21-1674-2.
- [Sil02] Aelian De Silva, *Technical Terms in Sinhala*, The Modern Book Company, 2002, ISBN 955-558-083-9.
- [Sor99] W. Sorata, *Sri Sumangala Sabdakosaya*, S. Godage and Brothers, 1999, Part 2.
- [Sor06] W. Sorata, *Sri Sumangala Sabdakosaya*, S. Godage and Brothers, 2006, Part 1.
- [Sta04] Sri Lanka Standard, *SLS 1134:2004 - Sinhala Character Code for Information Interchange*, Sri Lanka Standards Institute, 2004, Revision 2.
- [Wij82] Harischandra Wijayatunga, *Practical Sinhala Dictionary*, Ministry of Cultural Affairs, 1982, Vol 1.
- [Wij84] Harischandra Wijayatunga, *Practical Sinhala Dictionary*, Ministry of Cultural Affairs, 1984, Vol 2.
- [Wij05] Harischandra Wijayatunga, *Gunasena Great Sinhala Dictionary*, M. D. Gunasena & Company Limited, 2005, ISBN 955-21-1423-3.

## Appendix

.com	වණි	වාණිජ	වණික්	කොම	විකම
.net	දැල	ජාල	ජාලය		
.info	තතු	විනැව්	පුවත්	පවත	වග
.biz	වාණිජ	වෙළඳ	සකුණ	ව්‍යාපා	ව්‍යාප
.aero	ගගන	ගුවන්	අඹර		
.asia	ආසියා	ආසි			
.cat	කැට්	කැටලන්			
.coop	සමුප	සමාගම්	සමුහ	සහය	සමා
.jobs	රැකියා	කිස	රක්සා		
.mobi	ජංගම	සර	වල	සැරි	වලන
.museum	කටුගේ	කෞ	කෞතු	කටුගෙය	පැරැකි
.name	නෙම	නාම	සිය	ස්ව	නම
.pro	ප්‍රවීණ	නිපුණ	විද්	වියති	වෘත්තික
.tel	හමුව	ටෙලි	හමු		
.travel	සංචාර	වාරිකා	සැරිසර	සැරි	ගමන
.gov	රජය				
.edu	සිප්	සතර	සරස		ඉගැනු
.mil	රණ	යුධ	හමුදා	යුද	
.int	ජගත්	ලෝ	ලොව	දියන	දැතුරු

Table 2: Selected gTLDs

Test No.	Sinhala Domain Name	Description
1	http://සරසව්.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://සරසව්.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

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. ෂ ් ෂ ෂ ෂ. 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

Visually Identical Characters	Visually Identical Term(s)	Correct Term(s)	Visually Identical Term(s)	Correct Term(s)
බ - බ	බබාට-නමක්	බබාට-නමක්	බබාට-නමක්	බබාට-නමක්
ඩ - ඩ	පාඩම්	පාඩම්	පාඩම්	පාඩම්
ජ - ජ	අංජනම	අංජනම	අංනෙම	අංජනම
ප - ප	කැඩපන	කැඩපන	කැඩපත	කැඩපත
සා - සා	සාණ	සාණ	සෘණ	සාණ
ඔ - ඔ	කොළඹ	කොළඹ	කොළඹ	කොළඹ
භ - භ	අභස	අභස	අභස	අභස
භ - භ	ඉභන	ඉභන	ඉභත	ඉභත
ච - ච	සචන	සචන	සචන	සචන
ජ - ජ	සමාජය	සමාජය	සමාජය	සමාජය
ඵ - ඵ	ඵතෙර	ඵතෙර	ඵතෙර	ඵතෙර
ඵ - ඵ	ඵලකිරි	ඵලකිරි	ඵලකිරි	ඵලකිරි
ඵ - ඵ	කඵවර	කඵවර	කඵවර	කඵවර
ඩ - ඩ	කණ්ඩලම	කණ්ඩලම	කණ්ඩලම	කණ්ඩලම
ඩ - ඩ	අඩහැරය	අඩහැරය	අඩහැරය	අඩහැරය
ඩ - ඩ	කඩවර	කඩවර	කඩවර	කඩවර
ම - ම	අමරදේව	අමරදේව	අමරදේව	අමරදේව
ධ - ධ	සද්ධර්මය	සද්ධර්මය	සද්ධර්මය	සද්ධර්මය
බ - බ	ප්‍රබල	ප්‍රබල	ප්‍රබල	ප්‍රබල
ච - ච	නචනම	නචනම	නචනම	නචනම
ච - ච	චැටකොලු	චැටකොලු	චැටකොලු	චැටකොලු
ච - ච	චනචර	චනචර	චනචර	චනචර
කඩ - කඩ	නිරුජ්කඩනි	නිරුජ්කඩනි	නිරුජ්කඩනි	නිරුජ්කඩනි
ඔ - ඔ	ඔලු	ඔලු	ඔලු	ඔලු
ඩ - ඩ	ලඩකාව	ලඩකාව	ලඩකාව	ලඩකාව
ඩ - ඩ	පොඩ්ඩක්	පොඩ්ඩක්	පොඩ්ඩක්	පොඩ්ඩක්
ජී - ජී	අච්ච්ච්න	අච්ච්ච්න	අච්ච්ච්න	අච්ච්ච්න
ණි - ණි	පැරණි	පැරණි	පැරණි	පැරණි
චි - චි	අස්චි	අස්චි	අස්චි	අස්චි
චි - චි	සුච්ච්ච්න	සුච්ච්ච්න	සුච්ච්ච්න	සුච්ච්ච්න
චු - චු	ඇහැචු	ඇහැචු	ඇහැචු	ඇහැචු
චු - චු	පචුචා	පචුචා	පචුචා	පචුචා
චු - චු	සබ්බචුචු	සබ්බචුචු	සබ්බචුචු	සබ්බචුචු
කු - කු	කුරුච්ච්ච්	කුරුච්ච්ච්	කුරුච්ච්ච්	කුරුච්ච්ච්
චු - චු	පුනර්ච්ච්ච්	පුනර්ච්ච්ච්	පුනර්ච්ච්ච්	පුනර්ච්ච්ච්
චි - චි	ධරණිච්ච්ච්	ධරණිච්ච්ච්	ධරණිච්ච්ච්	ධරණිච්ච්ච්
චු - චු				

Table 6: Homoglyphs within the Sinhala script

Script	Glyph	Confusing Sinhala Glyph
Bengali	ঞ (U+098C)	ච (U+0DC0)
Gujarati	ૢ (U+0AB3)	ඡ (U+0D9C) (U+0DD4)
Gujarati	ૣ (U+0A9D)	ඣ (U+0DB1) (U+0DCA)
Telugu	ఱ (U+0C0E)	ఱ (U+0DB6)
Telugu	ఱ (U+0C28)	ఱ (U+0DB6) (U+0DCA)
Telugu	ఱ (U+0C2C)	ఱ (U+0DC3)
Telugu	ఱ (U+0C30)	ఱ (U+0DBB) (U+0DCA)
Telugu	ఱ (U+0C35)	ఱ (U+0DB6) (U+0DCA)
Kannada	ಱ (U+0C8E)	ಱ (U+0DB0)
Kannada	ಱ (U+0CAC)	ಱ (U+0DC3)
Malayalam	ഠ (U+0D15)	ഠ (U+0D9A)
Malayalam	ഠ (U+0D17)	ഠ (U+0D9C)
Malayalam	ഠ (U+0D28)	ഠ (U+0DC6)

Table 7: Homoglyphs in other scripts