Issues pertaining to rendering and resolving Labels with Conjunct Consonants, Rakaransaya and Yansaya forms in Sinhala Script

Harsha Wijayawardhana B.Sc. (Miami), CITP(UK), FBCS(UK) in collaboration with LK Domain Registry

Chair LLWG and COO and CTO Theekshana

1) Introduction

In 2010, Vint Cerf, who is considered one of the fathers of the Internet, unveiled two ccTLDs in their native scripts for Sri Lanka, dot Lanka (.ලංකා) in Sinhala Script and dot ilangai (.இலங்கை) in Tamil. Those two ccTLDs became some of the first non-Latin country top-level domains to be delegated by ICANN. Unusual though, " ශ්‍රී "(Shri) in Sinhala Script did not make it to the Country Code Top Level Domain. Ancients called Sri Lanka by many names. Greeks called it Thaprobane, and in Valmiki's Ramayana, Sri Lanka is identified as Lanka. Although "Sri" is a recent addition to the country's name, throughout Asia, Sri Lanka is known as Lanka for thousands of years. It is not history or culture that determined not adding Sri to Lanka in the ccTLD, but it was a technical issue of having a hidden character, Zero Width Joiner that bars "Sri" from being used for Top Level Domains. Sinhala “Sri” Glyph is made up of three parts, the letter Sha (ශ), Rakaransaya, which forms an arch at the bottom, and the vowel modifier long I on top of the letter, ශ. Rakaransaya and Yansaya come under special ligatures, which are known as Consonant Conjuncts. Rakaransaya, a special Ra (ර) with Halanta(Virama) or Hal Lankuna in Sinhala, forms an arch underneath a consonant. The other consonant conjunct is Yansaya, special ya (ය), again with Halanta or Hal Lakuna. Yansaya forms a ligature with a consonant ka (ක) kya (ක්‍ය).

2) Rendering of Rakaransaya, Yansaya and Repaya

Rakaransaya, Yansaya special forms under Consonant Conjuncts. Repaya is also a special Ra with Halanta (ර්), which replaces ර් of a word, placing the Repaya glyph on top of the consonant next to ර්: හර්ෂ to හර්ෂ. Rakar, Yansa, and Reph forms are not encoded in Unicode. The above three forms are stored as follows:

- Rakaransaya: U+0DCA (්) U+200D (Zero Width Joiner) U+0DBB (ර)
- Yansaya: U+0DCA(්) U+200D (Zero Width Joiner) U+0DBA (ය)
- Rephaya (්)U+0DBB (ර) U+0DCA (්) U+200D (Zero Width Joiner)

Zero Width Joiner (ZWJ) indicates to the Rendering Engine to form ligatures given above. In the Sinhala Conjunct form, ZWJ forms the ligatures binding two consonants together, removing the Virama or Halantha.

All web applications were used to strip ZWJ, fearing phishing attacks by hackers exploiting the presence of a hidden character in a text string in early 2000 with the release of IDNA2003. Removing ZWJ resulted in breaking up conjunct forms. Breaking up of ligatures or conjunct forms, which in Sinhala are called Bandi Akuru (බැඳී අකුරු), does not give an awkwardness due to both forms ---conjunct form without Virama or non-conjunct form with Virama--- being acceptable. Non-Conjunct forms with Virama are the norms after the Sri Lankan independence. However, the breaking up of Conjunct Consonants created an uproar among Sinhala scholars, forcing Government to take stern action to correct them. The Local Language Working Group (LLWG) made representations to Google and Microsoft, requesting them not to strip off the ZWJ. By 2007, most application developers, including Multinationals such as Google and Microsoft complied except in the address bar in their respective browsers.

IDNA 2003 protocol especially blocked using ZWJ or hidden characters and German and Cyrillic characters for Domain Registration, such as German Sharp S (ß) and Greek ending Sigma (ς). Owing to the barring of ZWJ in IDNA 2003, none of the words with ZWJ became candidates for ccTLD. However, in the second level, LK Domain Registry decided to issue both Sri (dot) Lanka without ZWJ and with ZWJ to the same applicant: ශ්‍රී. ලංකා and ශ්‍රී.ලංකා.

IDNA 2008 protocol replaced RFC 3490 and was released as RFC 5890. And RFC 5894 gave the authority to registrars to act cautiously to issue labels with ZWJ and ZWNJ instead of blocking them. The following is the wording in RFC 5894 on the labels with ZWJ and ZWNJ:

"A distinction is made between characters that indicate or prohibit joining and ones similar to them (known as CONTEXT-JOINER or CONTEXTJ) and other characters requiring contextual treatment (CONTEXT-OTHER or CONTEXTO). Only the former require full testing at lookup time"

"For example, a registry dealing with an Indic script that requires ZWJ and/or ZWNJ as part of the writing system is expected to understand where the characters have a visible effect and where they do not and to make registration rules accordingly."
(Excerpts from RFC 5894)

In IDNA 2008 protocol, those labels with Cyrillic and German Characters that were disallowed in IDNA 2003 were allowed for registration.

3) Implementation of IDNA 2008 protocol

Both popular browsers, such as Google Chrome and Opera, still do not support IDNA 2008 protocol. The hostname, ශ්‍රීලංකා.විශ්වස්ම්මුති.ලංකා, was checked on three browsers. It found that, while Google Chrome and Opera did not support IDNA 2008 protocol, the Firefox browser generated Punycode the label with ZWJ and complied with fully of the IDNA 2008 protocol.

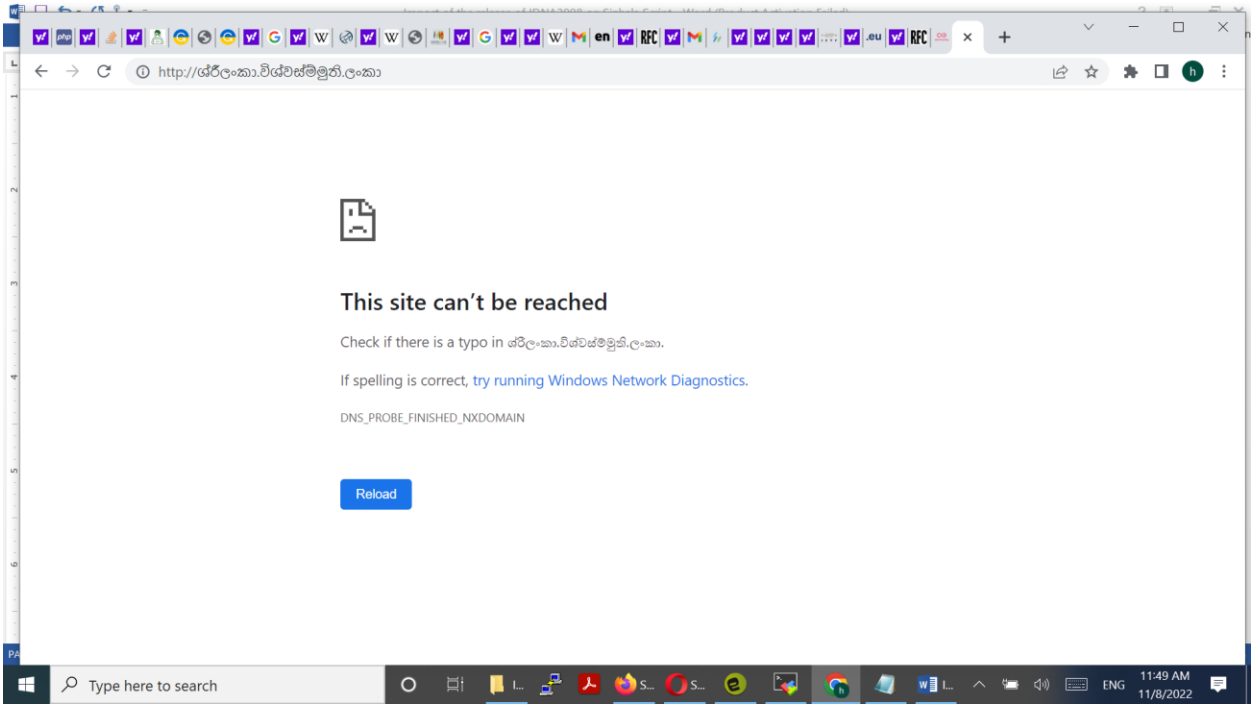

Figure 1Chrome Browser

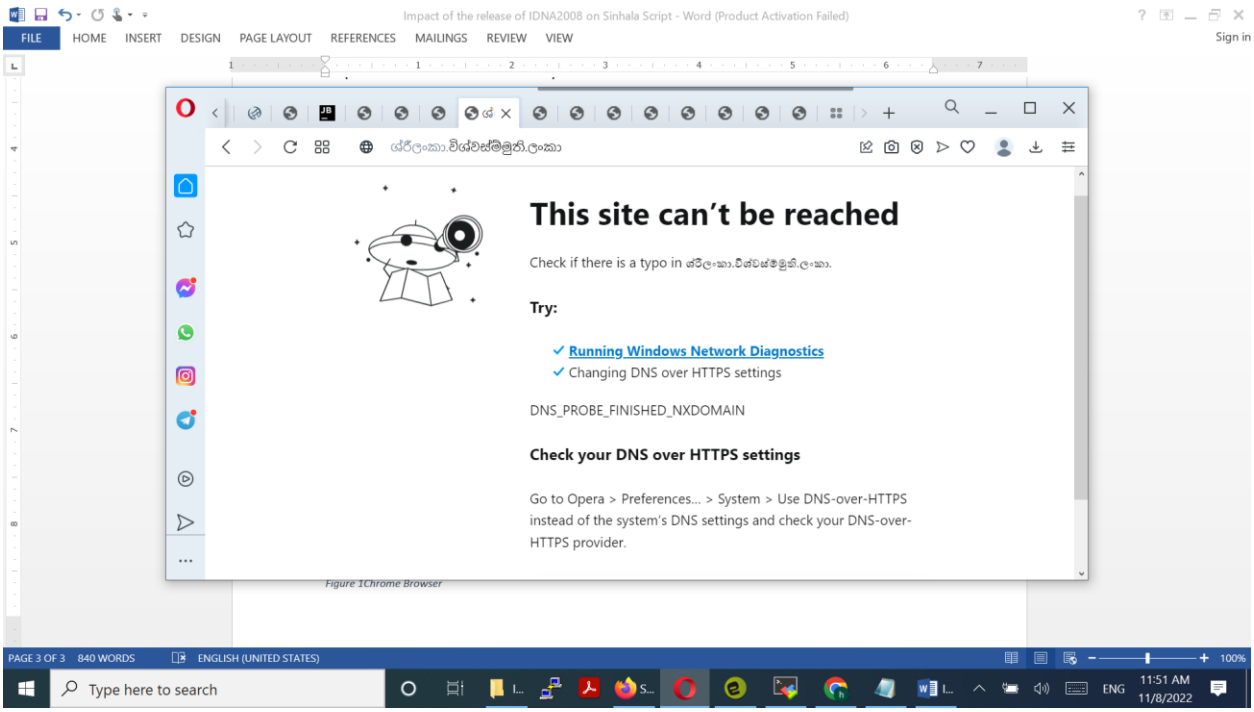

Figure 2 Opera

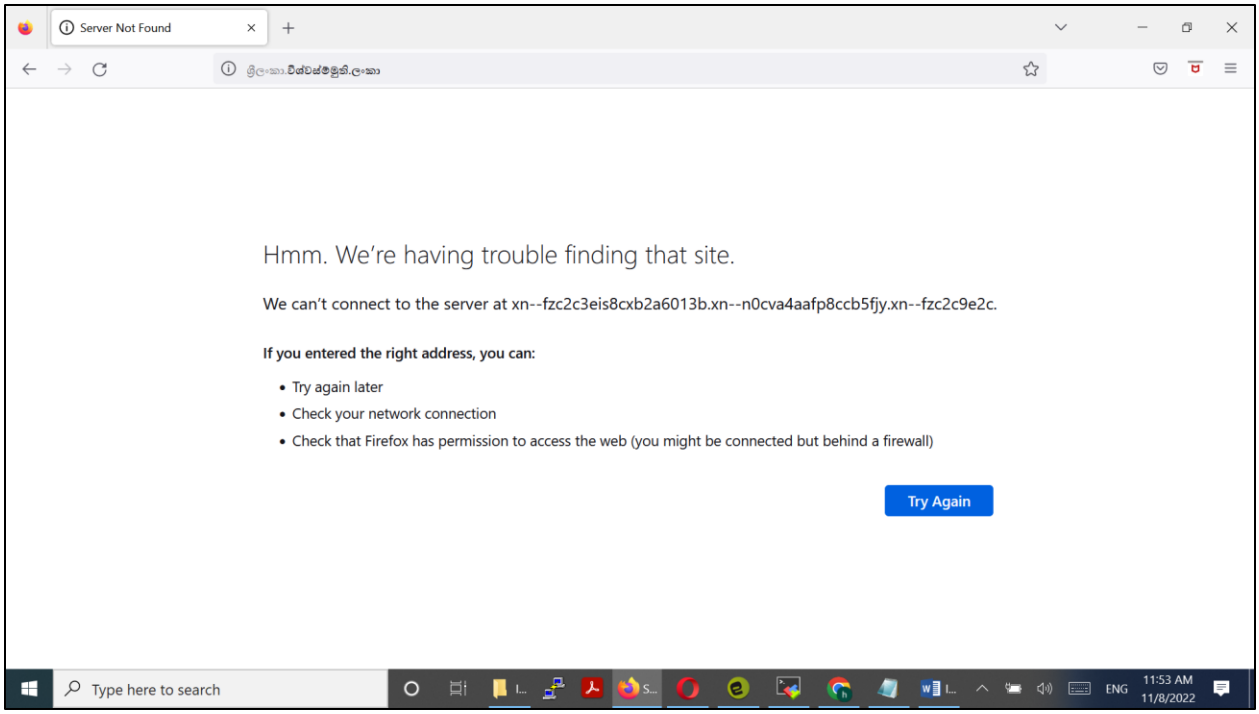

Figure 3 Firefox browser Please note Punycode generated

Punycode for ශ්‍රීලංකා.විශ්වස්ම්මුති.ලංකා: xn--fzc2c3eis8cxb2a6013b.xn--n0cva4aafp8ccb5fjy.xn--fzc2c9e2c

ශ්‍රීලංකා.විශ්වස්ම්මුති.ලංකා: xn--fzc2c3eis8cxb2a.xn--n0cva4aafp8ccb5fjy.xn--fzc2c9e2c

4) Conclusion

Most web applications are slow to support IDNA 2008 protocol. It was found that several Punycode generators online also do not support IDNA 2008 either. Sinhala Generation Panel has suggested blocking labels with ZWJ from registering as gTLD by disallowing those labels from the LGR Root Zone. Sinhala has many Sanskrit words, which have two Conjunct Consonants. Some of those words are very much like "Sri" candidates for either the second or third-level labels that cannot be overlooked and are much-desired possible labels.