Internationalized Domain Names -
Getting them to work

Gihan Dias
LK Domain Registry

What is IDN?

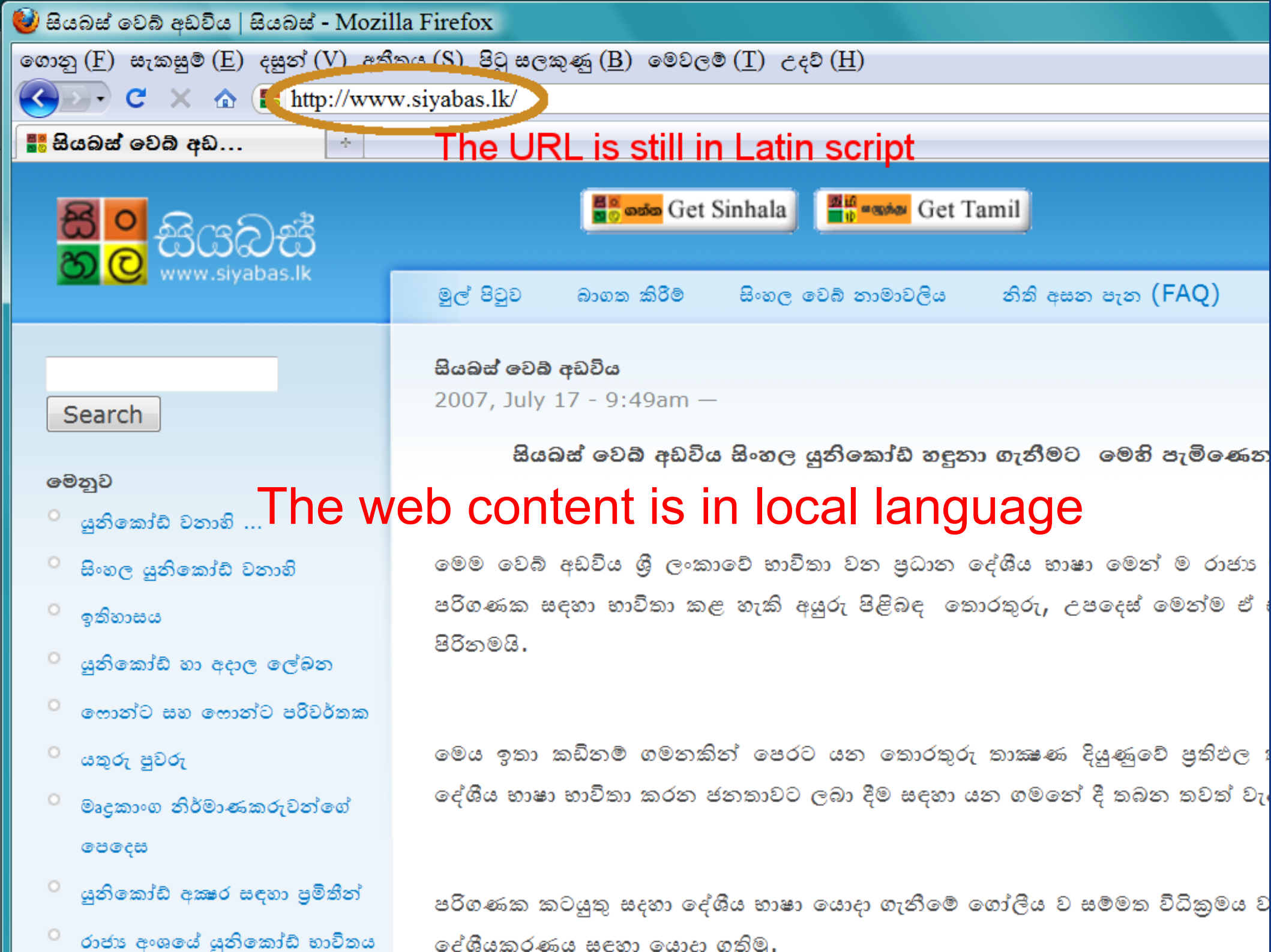
- Originally DNS names were restricted to the characters a-z (letters), 0-9 (digits) and '-' (hyphen) (LDH)
- Most languages use many other characters
  - accents and additional characters in Latin-based languages
  - non-Latin scripts
- IDN is a method of using domain names in other scripts
  - while maintaining compatibility with the ASCII DNS

IDN is…

- An abbreviation for “Internationalized Domain Name”
- A domain name where at least one character is outside the Letter-Digit-Hyphen set used in the original DNS
- Based on Unicode (ISO 10646) characters
- Major transition from 38 ASCII characters to thousands of Unicode characters

Why IDNs?

- Largest number of Internet users now in the Asia-Pacific
- Growing number of users not familiar with ASCII
  - or prefer to use their own script
- Using ASCII Domain names is significant linguistic barrier
- Native speakers of Arabic, Chinese, Japanese, Russian, Hindi, Tamil, Thai and others at considerable disadvantage
- Some names cannot be clearly expressed in Latin
- Need to provide a seamless experience
  - without toggling keyboards and typing unfamiliar characters


සියඹස් වෙබ් අඩවිය | සියඹස් - Mozilla Firefox
ගොනු (F) සැකසුම් (E) දසුන් (V) ඉතිහාසය (S) පිටු සලකුණු (B) මෙවලම් (T) උදව් (H)
http://www.siyabas.lk/
සියඹස් වෙබ් අඩ...
The URL is still in Latin script
සියඹස්
www.siyabas.lk
Get Sinhala
Get Tamil
මුල් පිටුව
බාගත කිරීම්
සිංහල වෙබ් නාමාවලිය
නිති අසන පැන (FAQ)
Search
මෙනුව
යුනිකෝඩ් වනාහි ...
සිංහල යුනිකෝඩ් වනාහි
ඉතිහාසය
යුනිකෝඩ් හා අදාල ලේඛන
ෆොන්ට සහ ෆොන්ට පරිවර්තක
යකුරු පුවරු
මෘදුකාංග නිර්මාණකරුවන්ගේ පෙදෙස
යුනිකෝඩ් අක්ෂර සඳහා ප්‍රමිතීන්
රාජ්‍ය අංශයේ යුනිකෝඩ් භාවිතය
සියඹස් වෙබ් අඩවිය
2007, July 17 - 9:49am —
සියඹස් වෙබ් අඩවිය සිංහල යුනිකෝඩ් හඳුනා ගැනීමට මෙහි පැමිණෙන
The web content is in local language
මෙම වෙබ් අඩවිය ශ්‍රී ලංකාවේ භාවිතා වන ප්‍රධාන දේශීය භාෂා මෙන් ම රාජ්‍ය
පරිගණක සඳහා භාවිතා කළ හැකි අයුරු පිළිබඳ තොරතුරු, උපදෙස් මෙන්ම ඒ
පිරිනමයි.
මෙය ඉතා කඩිනම් ගමනකින් පෙරට යන තොරතුරු තාක්ෂණ දියුණුවේ ප්‍රතිඵල
දේශීය භාෂා භාවිතා කරන ජනතාවට ලබා දීම සඳහා යන ගමනේ දී තබන නවත් වැ
පරිගණක කටයුතු සඳහා දේශීය භාෂා යොදා ගැනීමේ ගෝලීය ව සම්මත විධික්‍රමය ව
දේශීයකරණය සඳහා යොදා ගනිමු.

How IDN works

- Uses ASCII characters “on the wire” between clients and servers
  - no change to the original DNS protocol
  - used in thousands of servers
- URLs based on Unicode (ISO 10646) characters
  - possible to also use legacy character sets
- Applications (e.g. web browsers) convert the Unicode labels to ASCII
- Called Internationalizing Domain Names in Applications (IDNA)

Punicode

- A case folding and normalization process
- Unicode strings converted to ASCII Compatible Encoding (ACE) (and back) using Punicode algorithm
- Domain labels starting with “xn--” represent ACE encoded “internationalized” label
- One-to-one correspondence between (normalised) Unicode label and ACE (Punicode)

Examples of Punicode conversion

Unicode string	ACE string
ascii.com	ascii.com
日本語.jp	xn--wgv71a119e.jp
தமிழ்.in	xn--rlcus7b3d.in
bücher.de	xn--bcher-kva.de
සිංහලidn.lk	xn--idn-u4k9u8ai4i.lk

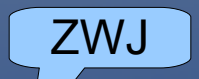
Issues in representing Indic languages with IDNA 2003

- A major issue is the handling of the zero-width joiner (ZWJ) and zero-width non-joiner (ZWNJ) characters
- These characters are used by most Indic scripts
- In some cases their use is optional (i.e. may change the appearance - but not the meaning - of a word)
  - In other cases they are required
- Omitting them will give an incorrect spelling or change the meaning

IDNA2003 Issues (cont.)

- IDNA2003 silently deletes ZWJ, ZWNJ and other control characters
- Results in some labels not displaying correctly

  e.g. ශ්‍රී → ශ්‍රී

  0DC1 0DCA 200D 0DBB 0DD3 → 0DC1 0DCA 0DBB 0DD3
- More than one Unicode label may map to the same Punycode label

Phishing issues

- Inter-script similarity
  - Certain letters in one script are similar (basically identical) to other letters in another script
  - e.g. Latin a, Cyrillic a
- Intra-script similarity
  - letter and numeral (or symbol) may look same

    e.g. Tamil அ (a) and Tamil அ (8)
  - two letters may look similar

    e.g. Sinhala ඩ and ධ

Browser issues

- Current browsers handle Unicode and automatically convert to Punicode
- May display punicode in URL bar instead of Unicode
  - anti-phishing feature
  - can be disabled via registry entry
- Some sites (e.g. Google) delete ZWJ/ZWNJ in search results
  - not an IDN issue, but annoying

Entering domain names

- Indic text is complex
- Some letters may need multiple keystrokes to enter
  - e.g. කෝ is entered using 4 keystrokes

IDN Process in Sri Lanka

- Sri Lanka formed an IDN task force
  - ICT Agency of Sri Lanka
  - Telecom Regulatory Commission
  - Official Languages Dept., Ministry of S & T, LK Domain Registry, ISPs, Universities, etc.
- First task – local lang. versions for .lk
- Identified potential strings
- Conducted Public Consultation

IDN Process (cont.)

- Obtained letters of support
- Submitted strings to ICANN
- Currently in evaluation process

IDN ccTLD Issues

- Labels for .lk in Sinhala and Tamil
  - short, meaningful ?, acceptable
- How many letters in a label?
  - what is a letter?
- Decided to use the full country name
- .ලංකා (.lanka) – Sinhala (4 keystrokes)
- .இலங்கை (.ilangei) – Tamil (6 keystrokes)

Selecting IDNs domains for organizations

- Names should be meaningful, easy to remember and easy to type
- Sinhala / Tamil names tend to be longer (in keystrokes) than English ones
- Abbreviations are less common in Sinhala/Tamil
  - some org. names sound strange when abbreviated
  - e.g. ICT Agency icta.lk => තොතාසනි.ලංකා
- Translate or Transliterate?

Selecting IDN strings for service names (e.g. www)

- www is a strange string
  - meaningless
  - difficult to pronounce
- How do we represent it in our language?
  - don't translate www
  - pick a string which clearly says “website”
- How about mail, rss, chat, proxy, etc.?

IDNA 2008

- Uses a different paradigm from IDNA 2003
- Most alphanumeric Unicode characters (except a few specified ones) are valid
- Applications (browsers) handle string mapping
- Registries handle registration restrictions
- Most differences between IDNA 2003 and IDNA 2008 are minor, except
  - IDNA 2008 allows ZWJ/ZWNJ in certain contexts
  - a few characters have different behaviour

IDNA 2008

- More information in tomorrow's tutorial

IDN Registration Policies

Registries (not only TLDs) need policies and procedures for:

- Handling of Indic numerals
- Unused and archaic letters
- Similar sounding letters
- Mixed-script labels
- Inter-label script mixing

LK Domain Registry Policies

- No registration of Indic numerals
- Registration limited to contemporary letters
  - should be able to register all company names, trademarks, ...
- No decision on similar-sounding letters
  - bundling?
- Allow registration of similar-looking letters
  - e.g. ඛ, ඛ ජ, ඡ etc.
- ZWJ only in specific contexts
  - more restrictive than IDNA 2008

LK Registry Policies (cont.)

- Allow Latin+Sinhala and Latin+Tamil mixed script labels
  - but no Sinhala+Tamil for now
- No inter-label script mixing
- Only register valid Unicode strings

LK Policies - Phishing

- No room for inter-script phishing
  - only allow Latin+Indic
  - no Latin letters similar to Indic
- Intra-script phishing possible
  - we don't register numerals or symbols
  - but some letters similar to others
  - e.g. (ඛ, ඛ; ච, ච; ඒ, ඒ)
  - users should be educated
  - bundling/reservation may be done

Domain Selection Guidelines

- No guidelines at present
- Sinhala/Tamil acronyms not common
- Recommend to select most significant word or words as domain name

Discussion