Internationalized Domain Names
Challenges and Solutions

Akshat S. Joshi
Centre for Development of Advanced Computing,
Pune, India
akshatj@cdac.in
New Delhi, 23rd Feb. 2012
IDNs in Indian perspective
The Multilingual diversity of India - Some facts & Figures

• The number of individual languages listed for India is 452. Of those, 438 are living languages.
• The Constitution recognizes 22 languages termed as Scheduled Languages.
• Four language families: Indo-Aryan, Dravidian, Tibeto-Burman, Munda.
• Two major script systems are used:
  • Brahmi - Left to Right writing system
  • Perso-Arabic - Right to left writing system
• 19 languages use 11 derivations of the Brahmi script.
• While 3 languages Urdu, Sindhi and Kashmiri uses the Perso-Arabic derivations.
<table>
<thead>
<tr>
<th>Sr. No</th>
<th>Family</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Indo-Aryan</td>
<td>• Assamese, Bengali, Dogri, Gujarati, Hindi, Kashmiri, Konkani, Maithili, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Sindhi, Urdu</td>
</tr>
<tr>
<td>2</td>
<td>Tibeto-Burman</td>
<td>• Bodo, Manipuri</td>
</tr>
<tr>
<td>3</td>
<td>Dravidian</td>
<td>• Kannada, Malayalam, Tamil, Telugu</td>
</tr>
<tr>
<td>4</td>
<td>Austro-Asiatic Munda</td>
<td>• Santali</td>
</tr>
</tbody>
</table>
The Multilingual diversity of India:

• One language - Many scripts
  • Santhali language gets written in
    • Devanagari script &
    • OlChiki script.
  • Sindhi language gets written in
    • Devanagari script &
    • Perso-Arabic Script

• One Script - Many languages:
  • Devanagari script is used for representing
    • Sanskrit, Hindi, Marathi, Nepali, Konkani, Maithili, Dogri, Bodo which are scheduled languages
Indian language complexities

**Syllable formation level**

ब्रह्मा

ब्र = ब + ् + र

ह्र = ह + ् + म + ि.

**Rendering order level**

किताब

किताब

(179+219+194+218+202)

**Alternate spellings**

अन्न

अन्न

बिरत

बिरट

**Alternate forms**

तस्वीर

तस्वीर

हिंदी

हिंदी

**Different inputting mechanism in Indian languages**
Most Indian languages are Multi-tier in nature.

When conjuncts come in picture, resulting glyph shapes increase manifolds.
IDN - Internationalized Domain Name

- Also called a multilingual domain.
- IDN is a domain name that contains characters from the Unicode character repertoire that other than letter/digit/hyphen (LDH) characters, which are Latin letters (a-z case ignored so includes A-Z), digits (0-9) and the hyphen (-).

Examples shown:

- Arabic: ﺍﻳﻜﺍ, ﻁﺭ@brief.com
- Chinese: 宜家.com, 上海酒店.com
- Greek: λλλ.com
- Hebrew: ﺳﺮﻴ,Yaron.com
- Japanese: バドミントン.com, [xn--0ouw9t.jp 江戸.jp]
- Korean: 한글.kr, 현금영수증.kr
- Russian: доменные-имена.com, IKEA.com
- Spanish: viña.delmar.cl, ñandú.cl
- Symbols: ®.com, ©.com
- Traditional Chinese: 台湾大学.tw, 中大.tw
IDN - Internationalized Domain Name

Also referred as multilingual domain name

• www.भारतभाषा.भारत (case of IDN.IDN in Hindi)

• www.भारतभाषा.com (case of IDN.ASCII in Hindi)

• www.भारतभाषा.भारत (case of IDN.IDN in Gujarati)
Punycode

- User enters IDN: www. य८६र०.com (Application such as browser converts to ASCII Compatible encoding (ACE)): www.xn--3b7vcv67.com
- Registry entry: xn--3b7vcv67.com (ASCII characters)

Vast amount of DNS support software only handles ASCII. Hence to accommodate characters from Unicode repertoire in domain names, Unicode to Punycode conversion is done.
Phishing is typically carried out by email spoofing or instant messaging, and it often directs users to enter details at a fake website whose look and feel are almost identical to the legitimate one.
Types of Variants:

1. Homographic variants: Similar Looking
   - 1 / l in Latin
   - त / त in Devanagari

1. Homophonic variants: Similar Sounding / alternate spelling
   - color / colour in Latin
   - हिंदी / हिंदी in Devanagari

1. Case variants:
   - C / c in Latin (No such case in Indian Languages)
Most of the browsers and applications using IDNs, display labels in minimal size.

This results in maximum number of spoofing and phishing attacks.

Multi-tier scripts such as used in Indian languages are less readable in the address bar.

Unicode normalization rules have also been considered as variants.
Homographic Variants:

- www.मुद्रा.भारत
- www.मुद्रा.भारत
- www.मुद्रा.भारत

\[ d + o + r = \text{dra} \]
\[ d + o + n = \text{dna} \]
\[ d + o + g = \text{dga} \]
## Homographic Variants:

### Telugu Variants:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Telugu Glyph" /></td>
<td><img src="image2" alt="Telugu Glyph" /></td>
<td><img src="image3" alt="Telugu Glyph" /></td>
</tr>
<tr>
<td>0CAC+0CCD+0CA6</td>
<td>0CAC+0CCD+0CA7</td>
<td>0CAC+0CCD+0CB2</td>
</tr>
<tr>
<td><img src="image4" alt="Telugu Glyph" /></td>
<td><img src="image5" alt="Telugu Glyph" /></td>
<td><img src="image6" alt="Telugu Glyph" /></td>
</tr>
<tr>
<td>0CA5+0CBF</td>
<td>0CA6+0CBF</td>
<td>0CA7+0CBF</td>
</tr>
</tbody>
</table>

### Tamil Variants:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image7" alt="Tamil Glyph" /></td>
<td><img src="image8" alt="Tamil Glyph" /></td>
</tr>
<tr>
<td>0B92+0BB3</td>
<td>0B94</td>
</tr>
</tbody>
</table>
Homophonic Variants & Alternate spellings:

- **Valid Homophones**: हिंदी versus हिन्दी
- **Common Misspellings**: इण्डिया versus इन्डिया

While formulating the IDN policy for .in we have not considered these variants as historically other domains have always considered alternate spellings of www.color.com and www.colour.com as separate entities.
Case Variants

• Case variants are not applicable in case of Indian Languages

• However Indian languages are rich in synonyms
Need for Variant Identification

- Invisible characters like ZWJ and ZWNJ can greatly amount to visual spoofing possibilities.
  - If permitted, their placement within the Domain Name/Label should be restricted to only most compulsory cases.

- In some cases, within the same script, two languages need different conjunct formation rules.

- Across the Operating systems, Rendering Engines and their versions, the rendering is not same.
Indian scripts introduce syllabic variants

श्व / श्थ

Such homographs need to be considered while identifying variants

श्र्व = श्व
0936 094D 0930 094D 0935

श्व = श्थ
0936 094D 0935
Need for Variant Identification : Display Aspect

ZWJ and ZWNJ :
Invisible characters like ZWJ and ZWNJ can greatly amount to visual spoofing possibilities. A clear decision needs to be taken regarding their inclusion in TLDs and if included, their placement within the Domain Name/Label.

Examples :

- Maharashtrā without ZWJ and ZWNJ
  Code Points: 092E 0939 O93E 0930 O93E 0937 O94D O91F O94D O930

- Maharashtrā with zero width joiner after हा

- Maharashtrā with zero width non-joiner after म
Need for checking well formed-ness of labels

Rendering Engine lacunae:

The well formed word "किताब" as seen in Address of Safari (Version - 3.1.2 (4525.22)) on MAC OS Version -10.4.11 (Tiger)

Actual display

Expected display  

Bidi Algorithms needed for Urdu, Sindhi and Kashmiri are more complex.
Need for checking well formed-ness of labels

Rendering Engine lacunae:
An ill-formed word composed of sequence
0915 + 093F + 094D + 0924 + 093E + 092C
as seen in IE (Version 8) on Windows XP

Actual display

Expected display किताब

Some applications are incapable of showing IDN labels and show punycode instead.
Indian IDNs in Browser Address Bar - rendering issues

Various Operating Systems and Browser combinations were considered during testing of IDN .in (.भारत)

In the URL, extra spaces are introduced after the Indian language domain text.

In the URL, font size of Indian language domain text is less; so proper rendering is not possible for Bengali text.
Implementation of IDNs in .in(.भारत) ccTLD

- A formalism based on ABNF has been put in place to validate desired domain name for each language based on syllabic structure.
- The applicable character sets for all official languages have been identified from the respective Unicode code charts for the script of the language.
- No intermixing of scripts is allowed
- Variant rules have been formalized for Domain Name label.
- Variants occurring syllables have been identified within each language.
- The variant set has been kept optimal ensuring safety of citizens without being too restrictive.
The proposed policy for IDN in Indian languages has following layout:

1. Generic rules in compliance with .in registry.

2. Rules specific to IDN in Indian languages.

3. ABNF formalism and language specific restriction rules to ensure a well formed Brahmi Syllable.

4. Language Tables which will ensure that only those characters pertinent within a given language are used.

5. Variant tables to reduce the risk of spoofing and phishing.
Following are the general policy guidelines in case of Indian language domain names:

- Only letters, digits, and hyphens will be permitted in a domain name. Names cannot begin or end with hyphens.
- Mixing of two scripts will not be permitted.
- Use of Zero Width Joiner/Zero Width Non Joiner will not be permitted.
- Language numerals and punctuations will not be permitted.
- Special symbols or stress markers will not be permitted.
Policy in brief

**ABNF Formalism**

To ensure that the correct formation of the Brahmi syllable is permissible, a study of the structure of the Brahmi syllable pertinent to the languages under survey has been undertaken and applied to the IDN’s. Since the Brahmi syllable can be best described in terms of a Backus-Naur formalism, this has resulted in an ABNF (Augmented Backus-Naur formalism) in which both digits as well as Hyphens are incorporated. ABNF ensures that a malformed syllable is not acceptable:

उददददत (which incidentally is Google typed using the Inscript keyboard and the site has been registered, as have been similar “transliterations of Microsoft, IBM)

**Restriction Rules**

Since ABNF is a generic formalism, restriction rules pertinent to each language or script will be defined. These rules will ensure that the ABNF formalism is clearly applied to a given language.

e.g. Hindi, all possible Nukta characters will be identified to ensure that no other character will be attached with a Nukta

**Valid Nukta Characters:** ॠ,ऌ,ॡ,ॢ,ॣ,।,॥,०

**Invalid Combinations handled by restriction Rules:**  Trafford + Nukta
consonant-syllable →
\*k(C[N]H) C[N] [H|D|B|X|BD|BX|M[D|B|X|BD|BX]]
| [CH]Z
| L[H|C[D|H|M[D]]]
| AC[D|X|M[D|X]]

vowel-syllable → V[D|B|X|BD|BX]

Syllable → consonant-syllable [Y] | vowel-syllable[Y]

IDN-Label → (Syllable | digit)*([dash](Syllable | digit))
Language wise restriction rules

Restriction rules are the additional filters which when applied to generic ABNF, it results to a Language specific ABNF. Languagewise restrictions are as follows:

**Hindi:**

1. Maximum permissible number of consonants to form a syllable up to 4 hence $k = 3$.
2. BD and BX combinations are Non-existent.
3. Nukta can be only allowed after following characters: 
   क (0915)  
   ख (0916)  
   ग (0917)  
   ज (091C)  
   ढ (0921)  
   ठ (0922)  
   फ (092B)
Example of Hindi ABNF

1. $H \mid D \mid B \mid X \mid M$ cannot occur in the beginning of an IDN domain name
   Example:

   ख
   किक
   कंक
   क़क

   As can be seen they will result automatically in a “golu” marking an invalid character. This is an intrinsic property of the Indic syllable and is quasi automatically applied.

2. H is not permitted after V, D, B, X, M, digit and dash
   Example:

   अ़
   क़े
   क़े
### Variant Tables

1. **HOMOGRAPH IDENTITY**

   To reduce spoofing and phishing attempts homographs or visual look-alikes within a pertinent script shall be identified. These homographs will constitute variant tables.

<table>
<thead>
<tr>
<th>U+0926 U+094D U+0917</th>
<th>हूँ</th>
<th>U+0926 U+094D U+0930</th>
<th>हूँ</th>
</tr>
</thead>
<tbody>
<tr>
<td>U+0926 U+094D U+0918</td>
<td>हूँ</td>
<td>U+0926 U+094D U+0927</td>
<td>हूँ</td>
</tr>
</tbody>
</table>

2. **FUNCTION OF THE VARIANT TABLE**

   The function of the variant table is to allow one of the homographs, debarring the other one. First use of either one of the characters shall automatically disallow the other in the case of a given word.

   e.g. If a user chooses समरुद्धी, the choice will automatically debar समरुध्दी, protecting against possible spoofing.
SPARING USE OF THE VARIANT TABLE
• Since exclusion tables based on variants can debar a large number of words commonly used, the variant table shall be used sparingly and only when absolutely necessary. Thus where two homographs have a very high frequency, they shall not be banned, since such a ban will not allow for registration of a large number of domain names.

VARIANT TABLES TO BE LIMITED TO LIGATURES ALONE
• Further the variant table shall apply only to ligatures or conjuncts or combination of two or more consonants and single characters that have homographic identity shall not be part of the variant table, the logic being that a native speaker can easily disambiguate single characters. It is the conjunct forms that can create confusions as shown in the earlier example.
• Thus घ ध shall not be permitted as variants although in the URL bar of the browser, they have a look-alike identity.
Normalization

• Since Unicode allows for two different ways of entering a given character, normalization shall be set in place.

• A good example of normalization is in Codepage 900: Devanagari where ड can be written as ड + or ड 095C

• It was decided that such normalization shall form part of the variant rules, subsequently they have been taken care of by IDNA 2008.

• As a safe-guard still, characters that are subject to normalize shall constitute part of the variant table.
The ABNF policy has been further reinforced by language-specific tables in which ONLY those characters which are pertinent shall be permitted. Thus Indian digits, symbols, rare characters such as the Sanskrit long vowels have not been admitted. This has resulted in a language specific chart in which only those characters which are pertinent have been enlisted.

Each language table has been vetted by experts in the respective language.

In addition a chart with classification as per the different entities defined in the ABNF is also appended for ease of use of the developer.
DETAILED LAYOUTS FOR LANGUAGES

DEVANAGARI: MARATHI

This proposal comprises the following Parts:

1. ABNF Formalism
2. Restriction Tables if any
3. Examples
4. The Code-chart for Marathi is compliant with the Unicode 5.1 Code-chart
   Characters marked in yellow will NOT be considered since they do not fall
   within the purview of the IDN policies as laid down.
5. The nomenclatural description of the above code-chart, in conformity with
   the above,
6. A variant table with the most important homographs being identified.
7. Experts/Bodies consulted.
Meeting the challenges

Translating the policy into action
1. Design of a Portal for the Registrars and Registrants

2. Design of a single Font

3. Floating keyboard for ease of entry

4. Peace of mind tools for ease of access

5. Validator for ABNF to ensure well-formededness of the syllable

6. Providing web-services
**Indian Domain Names Registration System (Beta Release)**

- **Inscript keyboard layout for Hindi language.**
- **Select font size:** Normal

<table>
<thead>
<tr>
<th>Key</th>
<th>Hindi Characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4 5 6 7 8 9 0 - ० १ २ ३ ४ ५ ६ ७ ८ ९ ० -</td>
<td></td>
</tr>
<tr>
<td>Tab</td>
<td>औ त आ इ ई उ ऊ ए ऐ ओ औ</td>
</tr>
<tr>
<td>Lock</td>
<td>ओ ए ई उ ऊ ए ऐ ओ औ न ण त थ द ध ज ढ झ ञ ट ठ ड ढ ण तृ थृ दृ धृ जृ ढृ झृ ञृ टृ ठृ डृ ढृ णृ तं न ब भ म य र ल व स य</td>
</tr>
<tr>
<td>Shift</td>
<td>र ल व स य</td>
</tr>
</tbody>
</table>

**Choose Language:**
- Assamese
- Bangla
- Gujarati
- Hindi
- Kannada
- Konkani
- Malayalam
- Marathi
- Nepali
- Oriya
- Punjabi
- Sanskrit
- Tamil
- Telugu

**Enter Hindi label:**

Enter only भारतभाषा if you want your URL to be www.भारतभाषा.भारत

**Validity Status:**
Valid Hindi Domain Name!

**PunyCode:**
xn--c2b2ahb7ad8beb8lb

**You may also want:**
राष्ट्रभाषा;

**Engine Response:**

**Punycode of Input Label**

**Variant possible for the given Label**

---

In compliance with IDN Policy Version 1.5
© C-DAC GIST Pune
Portal to demonstrate functionality of Language based Web Services over the internet.
For latest developments, visit:

idn.cdac.in
Nurturing living languages

धन्यवाद

Nurturing living languages

दीवित भाषाओं का पोषण