Internationalized Domain Names for Applications (IDNA)
Agenda

• Understanding basic Domain Name System terms
• About IDNA protocol
• IDNs in Indian Language Perspective
Domain Name System

• The DNS maps IP addresses into unique alphanumeric addresses called domain names

• A domain name is string of characters which is easy-to-remember “handle” to connect to a computer over internet

• Domain name is just a pointer to an IP address
Anatomy of Domain Name

• A domain name consists of multiple strings separated by “.”
• Each string separated by ‘.’ (dot) is called a “Label”
  e.g. www, example, com

```
www.example.com
```

```
+---+  +---+
| sub-domain | name |
|           +---+
|                TLD |
+---+  +---+
  domain name
```
Anatomy of Domain Name continued...

- Top Level Domain (TLD)
  - The rightmost part of domain name
  - Only delegated to responsible organizations (Registry)
  - Can be ccTLD (Country Code TLD) or gTLD (Generic TLD)
Anatomy of Domain Name continued...

• Country Code Top Level Domain (ccTLD)
  • Country specific domains
  • e.g. “.in” (India), “.jp” (Japan), “.cn” (China)

• Generic Top Level Domains (gTLD)
  • General purpose domains
  • e.g. “.com”, “.org”
Anatomy of Domain Name continued...

• Second Level Domain
  • the "name" that you register along with the TLD
  • e.g. “example.com” “wikipedia.org”
Anatomy of Domain Name continued...

• Sub Domain
  • Part of a larger domain. They are optional
  • Based on the user’s choice

• e.g.
  community.oracle.com
  cloud.oracle.com
  docs.oracle.com
Internationalized Domain Name (IDN)

• Domain name in your native language or script.
• Allow characters from different scripts, beyond the Letters (a to z), Digits (0 to 9) and Hyphen (-) : LDH

भारतभाषा.भारत (case of IDN.IDN in Hindi)
भारतभाषा.com (case of IDN.ASCII in Hindi)
भारतभाषा.भारत (case of IDN.IDN in Gujarati)
Need for IDN – Bridging the Gap

• Up till now domain names could be allowed in Latin characters only.
• IDNs open new horizons to have domain names in multiple languages or scripts.
• An opportunity for everyone in the world to be able to access the Internet in their native language
How IDN works in Domain Name System

• DNS can only understand Latin characters, hyphens and digits.

• Internationalized Domain Name is converted to Punycode (sequence of Latin characters, digits and hyphen) which is then given to the DNS for processing.
Where it sits?
IDNA Protocol – IDNA2003

• Introduced in 2003

• Used Unicode version 3.2

• The series of RFCs collectively known as IDNA
  • RFC 3490, 3491, 3492, 3454

• IDNA is defined for labels, not for parts of them and not for complete domain names
IDNA2003 Mechanism

- Transforming (mapping) a Unicode string to remove case and other variant differences.
  
  \[
  \text{rukam} \rightarrow \text{rukam} \\
  \text{U+0958 U+0926 U+092E} \rightarrow \text{U+0915 U+093C U+0926 U+092E}
  \]

- Checking the resulting string for validity, according to certain rules.
  - e.g. Bi-Directionality Rules

- Transforming the Unicode characters into a DNS-compatible ASCII string using a specialized encoding called *Punycode*
  
  \[
  \text{rukam} \rightarrow \text{xn--11b8aya8c} \\
  \text{U+0915 U+093C U+0926 U+092E} \rightarrow \text{xn--11b8aya8c}
  \]
Issues with IDNA2003

• IDNA 2003 is tied to Unicode 3.2

• Mapping Phase Issues:
  • What “can be used” instead of what “can be registered”

• Supporting tables for IDNA appears to be far more sensitive to subtle changes
IDNA Protocol – IDNA2008

• Removed dependency on any Unicode version
• Categorized Unicode Code points
  • PVALID, CONTEXTUAL RULE REQUIRED, DISALLOWED, and UNASSIGNED
• New set of RFCs defining the IDNA2008 protocol
  • RFC 5890, 5891, 5892, 5893, 5894
• RFC 3492 for “Punycode” conversion remains unchanged
• Only specifies what is allowed “to be registered” in the DNS old
  IDNA specifies what is allowed “to be used” which include
  mapping phase
PunyCode

• Transform a Unicode string uniquely and reversibly into a smaller, restricted character set.

• Process
  • Separation of ASCII characters
  • Encoding of non-ASCII character to ASCII sequence

"bücher" → "bcher-“ → “bcher-kva” → “xn--bcher-kva”
Challenges

• Phishing
  • Phishing is the way of obtaining user credentials like credit card details and other confidential details by pretending to be a trustworthy entity.

Actual Websites

- paypal.com
- sbicard.com

Phished Websites

- paypa1.com
- sbicards.com

Brand name abuse

- youtubç.com
- fašebook.com
- maiłgoogle.com
- microsoft.com
IDNs in Indian Language Perspective

• Complexity of Indian Languages pose several issues to IDN.
• The number of visual variations possible with Indian Languages are enormous
## Indian Language ccTLDs

<table>
<thead>
<tr>
<th>ccTLD</th>
<th>Punycode</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>भारत</td>
<td>xn--45br5cyl</td>
<td>Assamese</td>
</tr>
<tr>
<td>भारत</td>
<td>xn--45brj9c</td>
<td>Bangla</td>
</tr>
<tr>
<td>भारत</td>
<td>xn--h2brj9c</td>
<td>Hindi, Marathi, Konkani, Nepali, Bodo, Dogri, Maithili and Sindhi</td>
</tr>
<tr>
<td>भारतम्</td>
<td>xn--h2breg3eve</td>
<td>Sanskrit</td>
</tr>
<tr>
<td>भारोत</td>
<td>xn--h2brj9c8c</td>
<td>Santali</td>
</tr>
<tr>
<td>भारत</td>
<td>xn--gecrj9c</td>
<td>Gujarati</td>
</tr>
<tr>
<td>भारत</td>
<td>xn--s9brj9c</td>
<td>Panjabi (Gurmukhi)</td>
</tr>
<tr>
<td>पंजाब</td>
<td>xn--2scrj9c3d</td>
<td>Kannada</td>
</tr>
<tr>
<td>മലയാളം</td>
<td>xn--rvc1e0am3e</td>
<td>Malayalam</td>
</tr>
<tr>
<td>پہاڑہ</td>
<td>xn--mgbbh1a71e</td>
<td>Urdu</td>
</tr>
<tr>
<td>ଓଡ଼ିଆ</td>
<td>xn--3hcrj9c</td>
<td>Odia</td>
</tr>
<tr>
<td>சிங்கள்</td>
<td>xn--xkc2dl3a5ee0h</td>
<td>Tamil</td>
</tr>
<tr>
<td>తెలుగు</td>
<td>xn--fcrj9c3d</td>
<td>Telugu</td>
</tr>
</tbody>
</table>
Variants: Scope for Phishing

www.मुद्रा.भारत

dra

www.मुद्रा.भारत

dna

dga
Restrictions
Solution - Validation of IDN String

• A formalism based on ABNF has been put in place to validate desired domain name for each language based on syllabic structure.
Solution - Variants

• Variant Identification, Generation and Action:
References

• RFC 5890, 5891, 5892, 5893, 5894
  • https://tools.ietf.org/html/rfc5890
• RFC 4690 - https://www.ietf.org/rfc/rfc4690.txt
• http://www.cdac.in/index.aspx?id=mlc_gist_idn
• http://unicode.org/faq/idn.html
• http://www.unicode.org/reports/tr46/#IDNA2008
• https://en.wikipedia.org/wiki/Punycode
• http://nixi.in
thank you