Domain Names in Pakistani Languages

سرد حسین اور رابع سریمندی

نیشائی پیشور سی آف کمپیوئر اینترنیت سانفریست
Domain name

- Domain name is the address of the web page on which the content is located.
Internationalized Domain Name (IDN)

- Domain name or address of the web page in local language is called an IDN
- Based on the Unicode standard
Morning Session

- Introduction to the Unicode standard
- Introduction to Internationalized Domain Names
- Issues related to IDNs for Pakistani languages
Afternoon Session

- **Exercises and Recommendations**
  - Character status revision at script level
  - Resolving confusability of characters
  - Additional composed characters
  - Digits and Mixing
  - Single vs. multiple language tables
  - Character and Label separator
  - ccTLD string
  - gTLD translations

IDNs for Pakistani Languages
Background: Unicode

- Everything in computers is represented as numbers
- Initially ASCII encoding
  - A → 65
  - B → 66 …
- Only supported Latin script, primarily English
- Other encodings developed for other languages, but cumbersome to develop separate encoding for each language of the world
Unicode

• Thus effort started to develop Universal encoding UNIcode
• Unicode Consortium develops the Unicode standard
• Covers almost all writing systems in current use today
• First version *The Unicode Standard 1.0* published in 1991
• Current version *The Unicode Standard 5.1* published in April 2008
• Adopted by industry leaders as Apple, HP, IBM, Microsoft, etc.
• Supported in many platforms including Java, Linux and Microsoft Windows, etc.
• Supported by many internationalized applications including Open Office, Firefox, Thunderbird, Microsoft Office, etc.
Unicode

- European scripts
  - Latin, Greek, Cyrillic, Armenian, Georgian, IPA
- Bidirectional (Middle Eastern) scripts
  - Hebrew, Arabic, Syriac, Thaana
- Indic (Indian and Southeast Asian) scripts
  - Devanagari, Bengali, Gurmukhi, Gujarati, Oriya, Tamil, Telugu, Kannada, Malayalam, Sinhala, Thai, Lao, Khmer, Myanmar, Tibetan, Philippine
- East Asian scripts
  - Chinese (Han) characters, Japanese (Hiragana and Katakana), Korean (Hangul), Yi
Unicode

- Other modern scripts
  - Mongolian, Ethiopic, Cherokee, Canadian Aboriginal
- Historical scripts
  - Runic, Ogham, Old Italic, Gothic, Deseret
- Punctuation and symbols
  - Numerals, math symbols, scientific symbols, arrows, blocks, geometric shapes, Braille, musical notation, etc.
Characters Semantics

- The Unicode standard includes an extensive database that specifies a large number of character properties, including:
  - Name
  - Type (e.g., letter, digit, punctuation mark)
  - Decomposition
  - Case and case mappings (for cased letters)
  - Numeric value (for digits and numerals)
  - Combining class (for combining characters)
  - Cursive joining behavior
Unicode is SCRIPT based

• One code per character per script
  – To avoid duplication of same letter used by multiple languages
  – For example:
    The character code 06A9 ك is same in Urdu, Sindhi, Pashto, Punjabi, Farsi, etc.

• Different code blocks reserved for different scripts

• For Arabic script
  • 0600, 0601, …, 06FE, 06FF
  • 0750…077F

IDNs for Pakistani Languages
Unicode is the basis for Internationalized Domain Names
Domain Name System (DNS)

- Domain name is the address of a website in the internet space which is used to access its contents from another machine.

```
1. www.crulp.org
2. www.crulp.org
ISP

3. 192.168.0.1
4. 192.168.0.1
5. Requested Found / Not Found
6. Request Reply

www.crulp.org = 192.168.0.1
```

(IDNs for Pakistani Languages)
Need of IDNs

- Current DNS is based on 7-bit ASCII standard, only supporting abc…xyz, 012…89, and ‘-’
- Makes it difficult to access Internet for people who do not understand English or Latin script
- We cannot change the overall existing system as it can break the internet
- The solution is to add layer that works on top of existing system
- IDN implements a mechanism which supports domain name in any language which can be converted to ASCII format and use the existing internet framework
- Initial set of protocols defined in 2003, called IDNA2003
Internationalized Domain Name in Applications (IDNA)

- A layer that takes the address in local languages and converts that into ASCII format (using toASCII() )
- DNS continues to resolve ASCII format as usual
IDNA 200X

- Some Issues observed in the original IDNA2003
  - Protocol dependence on Unicode ver. 3.2
  - Hardcoded language specific separators
- Decision to revise the original standard taken in 2006
- New standard, IDNA 200X currently under development
IDNA 200X

- Assigns values to all Unicode Character Database (UCD) on the basis of Unicode Properties
  - PROTOCOL VALID (or allowed)
  - DISALLOWED
  - CONTEXTO or CONTEXTJ (depends on the context of use)
Morning Session

✓ Introduction to Unicode
✓ Internationalized domain names
• Issues related to IDNs for Pakistani languages
Arabic Script

- Arabic script is the second largest script after Latin script.
- It is used for writing Arabic, Urdu, Persian, Balochi, Pashto, Sindhi and many other languages across Pakistan and the world.
- Arabic script is defined from:
  - U+0600 to U+06FF
  - U+0750 to U+077F
  - U+FB50 to U+FDFF (Obsolete presentation forms)
  - U+FE70 to U+FEFF (Obsolete presentation forms except U+FDFx sequence)
  - New addition of dot-less characters and separate dots
Arabic Script

- Cursive script
  - Shape of each letter may have four different shapes depending on its position (isolated, initial, medial or final)
- Bidirectional
  - Letters written from right to left
  - Numerals written left to right
- Diacritics (optionally) used for vowels
- Stretched shapes used for text justification
- Shapes of letters highly context sensitive
**Contextual Shapes of Different Letters**

<table>
<thead>
<tr>
<th></th>
<th>Isolated</th>
<th>Initial</th>
<th>Medial</th>
<th>Final</th>
</tr>
</thead>
<tbody>
<tr>
<td>ب</td>
<td>با</td>
<td>كبا</td>
<td>كب</td>
<td></td>
</tr>
<tr>
<td>ج</td>
<td>جيا</td>
<td>كجا</td>
<td>كيج</td>
<td></td>
</tr>
<tr>
<td>و</td>
<td>NA</td>
<td>NA</td>
<td>جو</td>
<td></td>
</tr>
</tbody>
</table>
Issues in Arabic Script Encoding

- Character status revision at script level
- Resolving confusability of characters
- Additional composed characters
- Digits and Mixing
- Single vs. multiple language tables
- Label separator
- ccTLD string
- gTLD translations
Character Status Revision at Script Level

- Currently a formula using character properties determines which character is PVALID or DISALLOWED.
- Some PVALID characters not used by any language and should be DISALLOWED.
- ASIWG recommendations (Handout pg. 2)
  - Quranic marks
  - Formatting marks
- Do we agree for Pakistani languages?
Confusability

- Visually similar character shapes create confusion
- Confusion can be due to initial, medial, final or isolated forms
- Different cases of confusability
  - Shape confusability
    - Exact shape confusion
    - Similar shape confusion
  - Composition confusability
Exact Shape Confusion

- كل = ل + ك looks same as گل = ل + ک

- ی + ل + ج چلی = (06CC) looks same as چلی = (0649) ی + ل + ج

- یا = ا + (06CC) يا = ا + (064A)
Similar Shape Confusion

- Urdu character ی (06CC) and Pashto character ی (06CD)

- Sindhi ک (06AA) and Urdu ک (06A9)
  - کا vs. کا
Composition Confusability

- There are characters that can be typed in more than one way.
  - \(U+0622 (\ddot{i}) = U+0627 (\dot{i}) + U+0653 (\ddot{a})\)
  - Although they look similar to the user, they translate to different ASCII codes.
<table>
<thead>
<tr>
<th>Composed Form</th>
<th>Decomposed Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>U+0622 (١)</td>
<td>U+0627 (١) + U+0653</td>
</tr>
<tr>
<td>U+0623 (٢)</td>
<td>U+0627 (١) + U+0654</td>
</tr>
<tr>
<td>U+0624 (٣)</td>
<td>U+0648 (و) + U+0654</td>
</tr>
<tr>
<td>U+0625 (٤)</td>
<td>U+0627 (١) + U+0655</td>
</tr>
<tr>
<td>U+0626 (٥)</td>
<td>U+064A (٦) + U+0654</td>
</tr>
<tr>
<td>U+0675 (٧)</td>
<td>U+0627 (١) + U+0674</td>
</tr>
</tbody>
</table>
Solution and Problem

- **Solution**
  - Mapping for confusable shapes
    - For Urdu ی (0649) can be mapped to ی (06CC)
  - Normalization for composed forms

- **Problem**
  - Unicode does not provide mapping
    - Language dependent
  - Only partial normalization is provided in the Unicode standard onto pre-composed characters
    - Script dependent
Issues in Arabic Script Encoding

- Character status revision at script level
- Resolving confusability of characters
- Additional composed characters
- Digits and Mixing
- Character and Label separator
- Single vs. multiple language tables
- ccTLD string
- gTLD translations
### Digit sets in Arabic

<table>
<thead>
<tr>
<th>ASCII</th>
<th>ARABIC-INDIC</th>
<th>EXTENDED ARABIC-INDIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>U+0300</td>
<td>U+0660</td>
</tr>
<tr>
<td>1</td>
<td>U+0301</td>
<td>U+0661</td>
</tr>
<tr>
<td>2</td>
<td>U+0302</td>
<td>U+0662</td>
</tr>
<tr>
<td>3</td>
<td>U+0303</td>
<td>U+0663</td>
</tr>
<tr>
<td>4</td>
<td>U+0304</td>
<td>U+0664</td>
</tr>
<tr>
<td>5</td>
<td>U+0305</td>
<td>U+0665</td>
</tr>
<tr>
<td>6</td>
<td>U+0306</td>
<td>U+0666</td>
</tr>
<tr>
<td>7</td>
<td>U+0307</td>
<td>U+0667</td>
</tr>
<tr>
<td>8</td>
<td>U+0308</td>
<td>U+0668</td>
</tr>
<tr>
<td>9</td>
<td>U+0309</td>
<td>U+0669</td>
</tr>
</tbody>
</table>
Mixing Digit Cases

1. Two sets are mixed
   - www.اردو.com
   - www.123اردو.com
   - www.٢١اردو.com
   - www.٣٢١اردو.com

2. No mixing of digits
   - www.اردو.com
   - www.123اردو.com
   - www.٢١اردو.com
   - www.٣٢١اردو.com

IDNs for Pakistani Languages
Mixing Digits

- Mixing digits
  - A large number of domain names can be generated
  - Many of the labels generated are linguistically incorrect
  - Users may perceive mixed digit labels similar to non-mixed ones; potential for spoofing/confusion
- No mixing
  - Number of domain names limited
  - Some languages may require mixing for complete representation of words
Mixing Digits

- Two of these digit blocks used by Pakistani languages
  - ASCII and Extended Arabic-Indic
- Which set is required in IDNs by the language?
- Is mixing of both types of digits allowed?
Character Separator

- Need a character separator for proper shaping in Urdu
  - Words may assume wrong shapes without a separator e.g. دس دن will be displayed erroneously دسدن without a separator
- Space not allowed in domain names
- Zero Width Non Joiner (ZWNJ)
  - But users unfamiliar with it
  - Not available on conventional keyboards
- Any alternate Solution?
Label separator

- Pakistani languages use +06D4 (\-\) as label separator
- Standard ASCII names in DNS use 002E (\.) as separator
- Using dash for Pakistani languages
  - Pros: Keyboard switching not required
  - Cons: Mapping has to be standardized for web browsers and other applications
- Using dot
  - Pros: Part of the existing Internet standard; no mapping is needed
  - Cons: Keyboard switching required
- What should be label separator?
Keeping in view the issues discussed so far…

- Language tables can be constructed in two ways
  - One table for each Pakistani language
  - Single table for all languages
- Both have advantages and disadvantages
### Single Language Table

- All languages represented in one table
- Lists needed and not needed characters for all languages in single table
  - Easier to maintain
  - New languages can be added conveniently
- But, how to deal with additional confusability? May compromise complete language being expressed
Multiple Language Tables

- One table for each Pakistani language.
  - For e.g. Baluchi, Pashto, Punjabi, Saraiki, Sindhi, Torwali
  - List each language’s character-set separately
  - Confusability is limited and can be addressed without compromising language expression
- But, difficult to maintain
- And difficult to upgrade develop separate table for each of the 66+ languages of Pakistan
ccTLD String

- Candidate Country-Code Top-Level Domain string

- پاکستان
- پک
## gTLD Translations

<table>
<thead>
<tr>
<th>gTLD String</th>
<th>gTLD Abbrev.</th>
<th>Urdu</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARPA</td>
<td>arpa</td>
<td>انترنیتی</td>
</tr>
<tr>
<td>COMPANY</td>
<td>com</td>
<td>کمپنی</td>
</tr>
<tr>
<td>EDUCATION</td>
<td>edu</td>
<td>تعلیم</td>
</tr>
<tr>
<td>GOVERNMENT</td>
<td>gov</td>
<td>حکومت</td>
</tr>
<tr>
<td>MILITARY</td>
<td>mil</td>
<td>فوج</td>
</tr>
<tr>
<td>ORGANIZATION</td>
<td>org</td>
<td>ادارہ</td>
</tr>
<tr>
<td>INTERNATIONAL</td>
<td>int</td>
<td>عالمی</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>gTLD String</th>
<th>gTLD Abbrev.</th>
<th>Urdu</th>
</tr>
</thead>
<tbody>
<tr>
<td>NET</td>
<td>net</td>
<td>نیت</td>
</tr>
<tr>
<td>INFORMATION</td>
<td>info</td>
<td>اطلاعات</td>
</tr>
<tr>
<td>MEDIA</td>
<td>media</td>
<td>میڈیا</td>
</tr>
<tr>
<td>NAME</td>
<td>name</td>
<td>نام</td>
</tr>
<tr>
<td>BUSINESS</td>
<td>biz</td>
<td>کاروبار</td>
</tr>
<tr>
<td>AEROSPACE</td>
<td>aero</td>
<td>فضاانیات</td>
</tr>
<tr>
<td>PROFESSIONAL</td>
<td>pro</td>
<td>پروفیشنل</td>
</tr>
<tr>
<td>MUSEUM</td>
<td>museum</td>
<td>میوزیم</td>
</tr>
<tr>
<td>Employment</td>
<td>jobs</td>
<td>ملازمت</td>
</tr>
<tr>
<td>Related</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Travel agents/Airlines</td>
<td>travel</td>
<td>سیاحت</td>
</tr>
</tbody>
</table>