INTERNATIONALIZED DOMAIN NAMES-(IDN)

Inam Ullah,
Convener: Consortium for Language Advocacy & Networking (CLAN)
Member: Task Force on Arabic Internationalized Domain Names (TF-AIDN)
What are Internationalized Domain Names?

“Internationalized Domain Names (IDNs) are domain names represented by local language characters. Such domain names could contain letters or characters from non-ASCII scripts (for example, Arabic or Chinese)”
An internationalized top-level domain (IDN TLD) is a top-level domain with a label that contains at least one character that is not a basic Latin letter (a,b,...z).
Domain names in Scripts other than Latin
IDNs are Script-Based

- IDNs are Script-Based NOT Language-Based
- Each script may be used by many languages
- For example, Arabic script is used by
  - Arabic
  - Pashto
  - Persian
  - Sindhi
  - Urdu…just to mention a few
Major Scripts used by the languages of Asia, Europe and Africa
IDN TLDs introduced after 2009

What has changed with Top-Level Domains?

**DOMAIN NAME**
Latin letters a to z without accents or symbols.

2009

**INTERNATIONALIZED DOMAIN NAMES ADDED**
Characters from different scripts.

- .org
- .com
- .uk

- .arb
- .china
- .rf
ASCII is limited to English
The Domain Name System (DNS) has not historically supported labels in other languages
Other languages use letters and script outside this character set (a-z), (0-9) and (-)
To make the DNS more accessible to users around the world
The internet community has been working Internationalizing the Domain Names
IDNs are essentially made up of Unicode characters
Users prefer to use their native languages over English.

The combined population of China and Japan (1.5 billion) people – that’s 5 times more than the U.S.

Internet is global but you can't register a domain name and have URLs in your language.

IDN is the answer.
Internationalized Domain Names allow people around the world to access domain names in their local languages.
ICANN’s IDN Program

Undertaking 28 + Scripts

Dealing with 19+ Generation Panels

There is a variation in the status of the Generation Panels.
Generation Panels' varying status

Label Generation Rules (LGR)

Generation Panel Status

Seated
Active
Forming
Initiating

Maximal Starting Repertoire (MSR)
Generation Penals under IDN Program

- Arabic
- Armenian
- Chinese
- Cyrillic
- Ethiopic
- Greek
- Japanese
- Khmer
- Korean
- Lao
- Neo-Brahimi
- Thai
- Tibetan
IDN cc-TLDs

IDN Country Code Top-Level Domains

ccTLDs: 47/37*
Arabic Script

Used by languages geographically distributed in the countries and regions of:

1. Middle East
2. South and West Asia
3. North and Sub-Saharan Africa
4. South-East Asia
Essentials for IDNs

- A set of Unicode Characters (Code Points)
- A set of specified variants
Major Languages Using Arabic Script

- Arabic
- Urdu
- Persian
- Pashto
- Uyghur
- Malay
- Hausa
- Wolof
TF-AIDN: Arabic GP

- Creation and oversight by community based Middle East Strategy Working Group (MESWG)
  - 33 members from 18 countries
  - Experts from a variety of backgrounds
Defining the code point principles
Finalizing the code point repertoire
Defining the code point variant principles
Finalizing the code point variant sets
Analyzing allocatable vs. blocked variants
Finalizing code point variant types
Analyzing Whole Label Evaluation rules
Finalizing WLE rules
Documenting the process
Creating XML LGR for Arabic script LGR proposal
Finalizing the Documentation and XML LGR proposal for submission to ICANN
<table>
<thead>
<tr>
<th>Description</th>
<th>Number</th>
<th>Language / Example</th>
<th>Glyph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of code points considered</td>
<td>343</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Code points PVALID according to IDNA 2008</td>
<td>289</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Code points included in MSR-1</td>
<td>239</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Code points for which TF-AIDN found evidence of use</td>
<td>167</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specific code points shortlisted but excluded in IDNA 2008</td>
<td>1</td>
<td>Jawi / U+0662 –</td>
<td>٢</td>
</tr>
<tr>
<td>Specific code points shortlisted but excluded in MSR-1</td>
<td>5</td>
<td>Sindhi / U+06FD</td>
<td>ع</td>
</tr>
<tr>
<td>Code points where TF-AIDN found evidence of use but which are out of scope based on general MSR rules (Threatened or declining language, optional character)</td>
<td>32</td>
<td>Ormuri / U+076B, Kurdish / U+0692</td>
<td></td>
</tr>
<tr>
<td>Number of code points suggested for LGR by TF-AIDN</td>
<td>129</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Current Code Point Repertoire
<table>
<thead>
<tr>
<th>الجزائر</th>
<th>إيران</th>
<th>عمان</th>
<th>المغرب</th>
<th>السعودية</th>
<th>سودان</th>
<th>فلسطين</th>
<th>بارزان</th>
<th>باكستان</th>
<th>الأردن</th>
</tr>
</thead>
<tbody>
<tr>
<td>تونس</td>
<td>قطر</td>
<td>مصر</td>
<td>قطر</td>
<td>فلسطين</td>
<td>العراق</td>
<td>موقع</td>
<td>موقع</td>
<td>موقع</td>
<td>موقع</td>
</tr>
</tbody>
</table>
## Situation of Pakistani Languages

### Number of speakers of larger languages (Wikipedia)

<table>
<thead>
<tr>
<th>Language</th>
<th>2008 estimate</th>
<th>Areas of Predominance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Punjabi</td>
<td>76,367,360</td>
<td>Punjab</td>
</tr>
<tr>
<td>Pashto</td>
<td>26,692,890</td>
<td>KP</td>
</tr>
<tr>
<td>Sindhi</td>
<td>24,410,910</td>
<td>Rural Sindh</td>
</tr>
<tr>
<td>Saraiki</td>
<td>18,019,610</td>
<td>Punjab</td>
</tr>
<tr>
<td>Urdu</td>
<td>13,120,540</td>
<td>Urban Sindh and urban Pakistan</td>
</tr>
<tr>
<td>Balochi</td>
<td>6,204,540</td>
<td>Balochistan</td>
</tr>
</tbody>
</table>
Almost all Pakistani languages use Arabic Script except:

Gujarati  (uses its own script)
Language tables almost finalized:
  Urdu & Pashto

Other Language Tables in progress:

Sindhi, Punjabi, Balochi, Balti, Shina, Seraiki,
Many characters can be confused with others

Problem exists in ASCII as well
  - Digit “1” and lower-case “l”
  - Digit “0” and upper-case “O”

IDNs increasing the character collection
  - From 64 in ASCII
  - To tens of thousands in Unicode

This kind of confusion
  - create opportunities for user mistakes
  - and fraud