SaudiNIC’s Proposed Solution

Registry-level Multilingual Arabic Script IDN Registration

IGF Workshop, Sharm Elsheikh, Nov 15-18, 2009

Presented by:
Abdulaziz H. Al-Zoman
Director of SaudiNIC, CITC
Chairman of Steering Committee
Arabic Domain Name Pilot Project

Developed by SaudiNIC/CITC Staff:
Raed Al-Fayez
Mohammed Al-Hamed
AbdulRahman Al-Ghadir
AbdulAziz Al-Zoman

Content

- Arabic Script Major Issues
- Confusing Similar Characters
- Proposed solution
  - Characteristics
  - Language-level required tables
  - Language-level required tables
  - Language-level operation
- Conclusion
The 2nd most widely used alphabetic writing system in the world

Used by many languages such as:
- Persian, Urdu, Turkish, Kurdish, Pashto, Jawi, …
There are a number of groups of characters that have the same shapes (Homoglyph).

- eg. Kaf, Heh, Yeh, Alef, … groups

- There are a number of groups of characters that have the same shapes (Homoglyph).

- eg. Kaf, Heh, Yeh, Alef, … groups
Security issues (stability, trust,...) e.g. phishing
  – Some should be addressed at language level first

Input devices (keyboards) are based on languages

Not all Arabic-script languages are ready:
  – Not widely/commonly used
  – Language community are not ready
  – Hard to make decisions on behave of other language communities
  – Pressure to start with ready languages

Many problems have been escalated from the protocol to be handled by the registry (e.g. variants, bundling ..etc)

... and yet has to provide a simple and transparent registration services
Confusing Similar Characters
Valid Variants

- Assume there are 4 variants to letter (ه)

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>هدهد</td>
<td>هدهد</td>
<td>بدهد</td>
<td>بدهد</td>
</tr>
<tr>
<td>هدهد</td>
<td>هدهد</td>
<td>بدهد</td>
<td>بدهد</td>
</tr>
<tr>
<td>بدهد</td>
<td>بدهد</td>
<td>مدهد</td>
<td>مدهد</td>
</tr>
<tr>
<td>مدهد</td>
<td>مدهد</td>
<td>مدهد</td>
<td>مدهد</td>
</tr>
</tbody>
</table>

16 possible ways to write “هدهد”
Only 4 are confusingly similar (25%)

Confusing Similar Characters
Handling Domain Name Variants

- It is expected that some domains will have a large number of variants, e.g.:
  - There are 16,384 possible variants to write the domain “هيئة الاتصالات-تقنية المعلومات”

Q. How to know if a variant of a domain name has been registered?

Store all variants → Waste of Time and Resources

Store only a master key →✓
Characteristics of A Desired Solution

- Work for both ccTLDs and gTLDs
- Easy and fast to be deploy by any registry
- Extendable to allow for adding new languages as they become ready
- Simple and Transparent for end users
  - Do not annoy/confuse end users with technical/special
  - Regular users should be able to register whatever they can type using available keyboards
- Based on standardized (or agreeable) policies and procedures
  - documented on RFC-like or Best-Practice documents

General overview

User

Registry functions
(Whois + Register + Activate)

Registry

Group variant table + Language tables

Proposal

For each language:
1-Language table & 1-variant table

Unicode
**Language Table (LT)**
- A set of code points (Base characters) to be used by a registry for registering IDN domains in the corresponding language.
- LT can have Alphabetical, Numbers and Separators (Hyphens, Dots)

**Variant Table (VT)**
- A table that records all relations of the LT characters with other characters across the script.
- Each relation is defined depending on its similarity either:
  - Exact similarity: refers to identically look between base character and another character (e.g. exact match/mirror image).
  - Typo similarity: refers to almost look between base character other character (e.g. typo/style match).
- Consists of a list of records, each record contains:
  - Base character (from LT),
  - List of other characters (variants) with:
    - A set of positions of similarity [Beginning, Medial, Final, Isolated],
    - Relation type (Exact, Typo)

### Examples of Variants

<table>
<thead>
<tr>
<th>Base Character</th>
<th>Typo Variants</th>
<th>Exact Variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>0641 FEH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>06A7 QAF WITH DOT ABOVE</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
List all possible shapes for the basic character

Search for all its variants from the rest of the Arabic script

Then compare the basic character with its variants in all possible positions.

Find all similarity position(s).

Record the similarity (type & position)

A base character from LT

A variant character from script

Exact

Typo

Compare
It combines all VTs into one table that group all base characters with all relations across script.

Each variant list will be assigned to a unique group key (master key) that identify that group and will be used for generating the Master Key.

Language Table (one for every supported language)

- Users can only register domains using base characters from only one language table.

Group Variant Table (GVT):

- Generated from variant tables.
  - It combines all VTs into one table that group all base characters with all relations across script.
  - Each variant list will be assigned to a unique group key (master key) that identify that group and will be used for generating the Master Key.
GVT keys

Keys are used for Querying GVT Relations for variant characters

Keys are used for Querying GVT
Registry-level Operation
Generating Master Key for a Label

1. Check if the input string follows certain language (using LT).

2. Generate UNICODE code for that input.

3. Identify the position for each character depending on language properties (UNICODE Standard).

4. Generate Master key by taking every code from (step 3) and do simple lookup in GVT.

Registry-level Operation
Finding Exact Variants

- Find all Exact strings using Master Key for activation purpose.

Master Key

G15B, G13M, G13I

Use Master key to reversal lookup and find all possible characters (Query relation)
Registry-level Operation

**Registrait Interface**

- Lookup process (whois)
  - Check domain syntax under any supported language using LTs.
  - Check if the same domain is available or not.
  - If it is found return the unavailable/whois-information; otherwise continue
  - Get the master-key for the domain (based on GVT)
  - Check if the master-key was registered before or not
  - If master-key is found return unavailable/whois-information; otherwise return domain is available

- Registration process:
  - Registrant should select one of supported languages and a domain (U-Label)
  - Registry should accept inputs based on the selected language table
  - If domain name can be registered (available based on Lookup process) then register the domain

- Activation process (enable exact variants)
  - Original Registrant can activate any exact variant from the registered domain’s Master Key.
  - List possible Exact variants that can be typed using one of the LT without intermingling between them
  - Activate one/many of Exact variants (if not activated before)

---

**Adding New Languages**

1. **Scan GVT keys in new GVT and check if keys with Q exist in any key in old GVT; if so take variant list of that key (from new GVT) and add it with variant list of old GVT.**

2. **Add the rest of key of new GVT at the end of old GVT keys.**

3. **Check the new GVT: if the keys with Q appear in different GVT keys or not**

*Failed Merge*

- Keys appear in different GVT keys

*Successful Merge*

- Keys don’t appear in different GVT keys

*Cure:*

- Regenerate old GVT using existing VTs including the new VTs.
- Then regenerate all old Master keys using new GVT!
We tried to have a prototype that fulfill the concepts of script based registry that is:
- Optimized, Simple, Transparent, Automated, and addresses many local issues

Next steps:
- Finalize the Language tables & variant tables for the Arabic Language.
- RFC or best-practice document.
- Variant TLDs should be delegated as part of ICANN’s FT
  - E.g. Arabic => کویرت => U+0643 U+0648 U+064A U+062A
  - Persian => کویرت => U+06A9 U+0648 U+06CC U+062A

Thank you!

Developing Team
Abdulaziz Al-Zoman, Raed Al-Fayez, Mohammed Al-Hamed, AbdulRahman Al-Ghadir

Thank you