SaudiNIC
Variant Management System

Raed Alfayez
SaudiNIC, CITC

1-Apr-2015, MENOG15
Agenda

- Introduction

- Variant Management System
  - Concepts
  - Requirements

- SaudiNIC’s VMS
  - Master Key Algorithm
  - Variants Filters
There are 64 “variants” for “Google.com” domain due to lower/upper case of ASCII letters.

- If you type any of them you will reach the same site
- The solution was done by DNS protocols
- All are allocated and delegated

But this is not the case for other languages!

- Arabic (کلی) vs. Urdu (کلی)!
The **2nd** most widely used alphabetic writing system in the world

- **Used by many languages** such as:
  - Arabic, Urdu, Persian, Turkish, Kurdish, Pashto, …etc

- **It is widely used by more than 43 countries**
  - more than one billion potential users could be concerned in using Arabic script domain names.

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

- eg. Kaf, Heh, Yeh, Alef, … groups
There is a need for a system to **solve and manage variants** in the whole Arabic script.

- It should be achieved through **coordination** between a Registry and Language communities.
- The goal is to **secure** the TLD name space in a simple and logical manner.
  - Enhance security (Limit domain phishing)
  - Ensure domain name reachability
  - Easy to use and manage
Variant Management System

Concepts and Requirements

Concepts

- One key for all variants
- Variants based on character position
- Single user input device
- International (cross-languages) reachability
- Simple user interface
- Study the whole script

Requirements

- Define Supported Language(s)
- Coordinate with Language Communities
- Solve any variant conflicts
- Finalize Variant Tables
- Use mechanism to group variants under one key
- Provide clever tools to manage variants
One Key for all Variants (Master key)

- Storing all possible variants is not a visible nor a practical solution, especially for longer domain names as they generate larger variant list.

- A new identification mechanism is needed to:
  - Easily manage the whole variants list with one unique identifier
  - Speed up the lookup process
  - Eliminate the need of saving all possible variants (save storage space)

- Example: Master Key algorithm

- Result:
  - Efficacy (Store only one Key instead of all possible variants).

<table>
<thead>
<tr>
<th>Label</th>
<th>Approximately # of variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>اتصال</td>
<td>300</td>
</tr>
<tr>
<td>اتصالات</td>
<td>6,000</td>
</tr>
<tr>
<td>الاتصالات</td>
<td>60,000</td>
</tr>
<tr>
<td>هيئة-الاتصالات</td>
<td>2,879,999</td>
</tr>
<tr>
<td>هيئة-الاتصالات-تقنية-المعلومات</td>
<td>82,944,000,000</td>
</tr>
</tbody>
</table>
Variants base on character position:

- In Arabic script languages, characters may take different shapes depending on their position (isolated, final, medial or initial) within a word.

- Therefore, a Variants Management System should consider character position when deciding if 2 code points are variants or not.

- Example (هدهد):
  - 0647 = 06BE = 06C1 = 06D5
  - 16 possible variants
  - 4 valid variants (25%)
  - 12 without risk (75%)

- Result:
  - More Accuracy
A label is composed using a single input character set table

- Arabic label can be typed using “one” keyboard layout (input device)
- There are no mixing between code points from different keyboards (Arabic Keyboard layout, Urdu keyboard layout ..etc).
- Example (کلی):
  - 18 Possible variants
  - 10 Blocked because of language mixing (56%)
- Example (القرآن-الكريم)
  - 11,900 Possible variants
  - 11,888 Blocked because of language mixing (99%)
- Result: More Accuracy (only valid allocate-able variants)
Visit our website: كلى.موقع

کلی.موقع

کلی.موقع

کلی.موقع

کلی.موقع

کلی.موقع

کلی.موقع

کلی.موقع

کلی.موقع

کلی.موقع
International Reachability

- End users should be able to reach their domain names regardless of their location.
- Input devices (language table) that would be used to reach a domain name (based on the user location) should be carefully considered when defining variants.
- For example:
  - A user registered the domain “كلى” (all characters from the Arabic language)
  - if another user try to reach that domain name from an Internet café in Pakistan he/she will type “كلى” (all characters from the Urdu language)
  - If that variant was not allocated, delegated and hosted then the domain name will not be reachable.
- In summary, variants need to be studied from both:
  - Similarity point of view (by language community) and
  - Reachability point of view (based on input devices used by other language communities).
Simple User Interface

- **Myths about registrants and variants:**
  - Registrant can decide which variant to allocate from a huge list of variants.
  - Registrant may know the differences between code points
    - Arabic KAF (U+0643) and KEHEH (U+06A9)
  - Registrant may know which variant should be allocated in order for their domain name to be reached globally.
  - Registrant can handle complex user interface to manage variants.
- **Please note: it is unpractical to list all allocate-able variants**
  - As the list may contain hundreds of allocate-able variants.
- **Hence: the Registry should help registrants to:**
  - Generate the best (desired) allocate-able variants (as helping examples)
  - Provide easy way to enable/disable them
  - Provide a way for advance users to manually type other desired allocate-able variants.
- **Registries should provide a separate web interface and an EPP command that list the best (desired) allocate-able variants using a clever way to help in managing variants**
  - i.e. using multiple filters to minimize the generated list (Variant Filters)
Variant Management System
Concepts (6): Study the whole Script

- Study variants across the **whole Arabic script**
  - A full study should be conducted across the whole Arabic Script in order to identify all possible variants against code points in the supported language table
    - Some existing solutions only check the variants between code points within only the support language tables.

  - This way whenever a **new language** is added there will be no need to **restudy** the previous supported languages and change their variant tables.

  - Result:
    - **less key regeneration** when adding new languages to the registry.
The Registry need to **choose** which language(s) will be supported under their TLD

Registry should **coordinate** with language communities (or language experts) to achieve the following:

- **Language Table,**
- **Variant Table (including Variant Types/Action)**
  - Language communities should study their code points across the whole script when identifying variants
  - Action on the variants (Allocated, Blocked ..etc)
- **Identify how their users may type a domain name using input devices from other languages**
  - Must be allocated variants
  - E.g. Arabic user may use Urdu keyboard to register and/or reach an Arabic domain name
Registry should **solve** any **conflicts** in variants or variants types

- **Examples:**
  - Language 1: A = B, Language 2: A <> B!
  - Language 3: C = D (Blocked), Language 4: C = D (Allocated)!

Registry **Finalize** the following:

- **Supported language tables(s)**
  - List of code points for each language
  - Will be used to stop **mixing** characters from different languages
- **Variant Table**
  - Variants, Variants Types/Actions (e.g. Allocate-able, Blocked)
  - Will be used to **generate** allocate-able variants
- **Filters**
  - Will be used to suggest **desired** variants

Registry should use a **mechanism to secure variants** from being registered by others by grouping them under one key

- E.g. Master Key algorithm

Registry should **provide a simple and clever** way for Registrants to:

- Register a domain name in their language
- List allocate-able and/or desired variants
- Enable and Disable allocate-able variants
SaudiNIC’s VMS

- SaudiNIC has developed a complete VMS:
  - Based on the stated Concepts
  - Provides the stated Requirements

- We developed a Master Key algorithm and Variant Filters to:
  - Secure our name space
  - Ensure domain name reachability
  - Simplify variants management
Generates a **unique key** for a domain name label and all of its possible variants (based on the character position), the new key can be used in the **lookup process** for both:
- Domain name availability
- Variants generation and allocation

For example:
- “G41B G42M G43F” represents 18 variants:
  - كلى (U+0643) (U+0644) (U+0649)
  - كلى (U+06A9) (U+0644) (U+0649)
  - كلى (U+06A9) (U+0644) (U+06CC)
  - كلى (U+06A9) (U+0644) (U+064A)
  - كلى (U+06A9) (U+0644) (U+064A)
  - كلى (U+06A9) (U+0644) (U+06CD)
  - كلى (U+06A9) (U+0644) (U+06D2)
  - … etc,

the full list: [http://arabic-domains.org/adn_tools/mk/index.php?T=1&M=%D9%83%D9%84%D9%89](http://arabic-domains.org/adn_tools/mk/index.php?T=1&M=%D9%83%D9%84%D9%89)
SaudiNIC’s VMS
Variant Filters

- **Goal:**
  - To reduce the huge size of allocate-able variants by intelligently display only the desired variants

- **How?**
  - Linguistically we study words in the Arabic language to find some rules to help identifying desired variants
  - We used N-grams model to statically study the repetitive patterns in Arabic words
    - An example of 2-gram for word “cars”: ca, ar, rs
    - We studied 2, 3 and 4-grams for more than 7 million non-repetitive words in the Arabic language
      - Source: Books, Newspapers, Refereed Academic Journals.. Etc.
  - We studied high-frequency patterns and then built some rules/filters based on them: (ال-، ال-، ال-، ال-،...) etc.
  - Then we developed a ranking system to order allocate-able variants based on weight given by each rule.
  - We have confirmed our findings with linguists and researchers.
Sample of our variant rules (21+ rules):

- **AlefMadaEnd**
  - Input: خطأ-ظلماء
  - Filtered: خطأ-ظلماء, خطأ-ظلماء, خطأ-ظلماء ..etc

- **AlefHamzaDownEnd**
  - Input: خطأ-ظلماء
  - Filtered: خطأ-ظلماء, خطأ-ظلماء, خطأ-ظلماء ..etc

- **Alf-Altareef:**
  - Input: القرآن
  - Filtered: القرآن, القرآن, القرآن

- **Alef-letter-Alef**
  - Input: رأيأت, رأيأت, رأيأت ..etc.
## SaudiNIC's VMS

### Variant Filters

#### II. Must be Al IV Not Desired Variants (252)

![Image showing filter options and results](image)

**Results:**

**Master Key:** G14I G42B G26M G35M G14F G14I G42B G36M G43M G14F

**Statistics Summary:**

<table>
<thead>
<tr>
<th>Category</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Variants</td>
<td>9999</td>
</tr>
<tr>
<td>I. Must be Allocated Variants</td>
<td>0</td>
</tr>
<tr>
<td>II. Desired Variants</td>
<td>3</td>
</tr>
<tr>
<td>III. Not desired Variants</td>
<td>252</td>
</tr>
<tr>
<td>IV. Blocked Variants</td>
<td>9744</td>
</tr>
</tbody>
</table>

#### I. Input:

<table>
<thead>
<tr>
<th>LANGUAGE</th>
<th>UNICODE</th>
<th>LABEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic, Persian, Urdu, Malay, Pashto</td>
<td>(U+0627) (U+0644) (U+062E)</td>
<td>الخطا-الظاهر</td>
</tr>
<tr>
<td></td>
<td>(U+0637) (U+0627) (U+002D)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(U+0627) (U+0644) (U+0638) (U+0645) (U+0627)</td>
<td></td>
</tr>
</tbody>
</table>
Easy interface for registrants:
More Information

- For more information about the **Master Key algorithm and Variant filters**:
  - [http://arabic-domains.org/docs/Master_Key_Algorithm.pdf](http://arabic-domains.org/docs/Master_Key_Algorithm.pdf)
  - Note: Will be updated to reflect new enhancements

- **Best practices** for managing Arabic domain names registries
  - Available soon
Thank you

للمزيد من المعلومات يمكنكم زيارة:
For more information you can visit:

سجل.السعودية

nic.sa

هيئة الاتصالات وتقنية المعلومات
Communications and Information Technology Commission

citic.gov.sa