



# ***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

# Introduction

- 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 (كلى) vs. Urdu (کلی)!

## Example of ASCII Variants

Google.com  
gOogle.com  
goOgle.com  
gooGle.com  
GooGle.com  
GoogLe.com  
...etc.

كلى

input[0] = U+0643  
input[1] = U+0644  
input[2] = U+0649

کلی

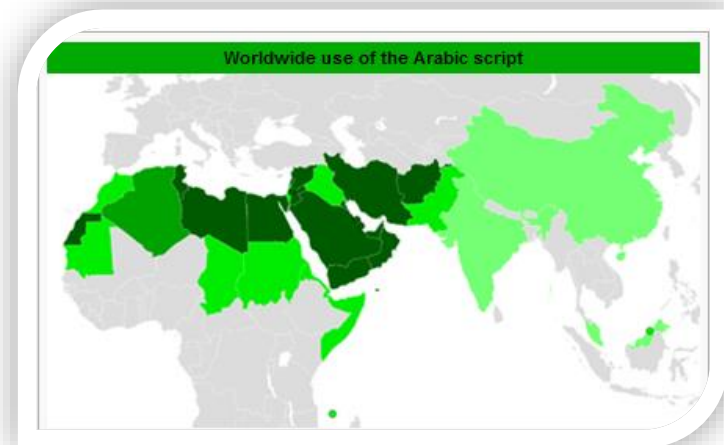
input[0] = U+06a9  
input[1] = U+0644  
input[2] = U+06cc



# Introduction

## Arabic Script

- The **2<sup>nd</sup>** 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.



Source: [http://en.wikipedia.org/wiki/Arabic\\_script](http://en.wikipedia.org/wiki/Arabic_script)

# Introduction

## Confusing Similar Characters



- There are a number of **groups** of characters that have the **same shapes** (**Homoglyph**).
  - eg. Kaf, Heh, Yeh, Alef, ... groups

	0600	0601	0602	0603	0604	0605	0606	0607	0608	0609	060A	060B	060C	060D	060E	060F
0	0600	0610		0630	0640	0650	0660	0670	0680	0690	06A0	06B0	06C0	06D0	06E0	06F0
1	0601	0611	0621	0631	0641	0651	0661	0671	0681	0691	06A1	06B1	06C1	06D1	06E1	06F1
2	0602	0612	0622	0632	0642	0652	0662	0672	0682	0692	06A2	06B2	06C2	06D2	06E2	06F2
3	0603	0613	0623	0633	0643	0653	0663	0673	0683	0693	06A3	06B3	06C3	06D3	06E3	06F3
4		0614	0624	0634	0644	0654	0664	0674	0684	0694	06A4	06B4	06C4	06D4	06E4	06F4
5		0615	0625	0635	0645	0655	0665	0675	0685	0695	06A5	06B5	06C5	06D5	06E5	06F5
6	0606	0616	0626	0636	0646	0656	0666	0676	0686	0696	06A6	06B6	06C6	06D6	06E6	06F6
7	0607	0617	0627	0637	0647	0657	0667	0677	0687	0697	06A7	06B7	06C7	06D7	06E7	06F7
8	0608	0618	0628	0638	0648	0658	0668	0678	0688	0698	06A8	06B8	06C8	06D8	06E8	06F8
9	0609	0619	0629	0639	0649	0659	0669	0679	0689	0699	06A9	06B9	06C9	06D9	06E9	06F9
A	060A	061A	062A	063A	064A	065A	066A	067A	068A	069A	06AA	06BA	06CA	06DA	06EA	06FA
B	060B	061B	062B	063B	064B	065B	066B	067B	068B	069B	06AB	06BB	06CB	06DB	06EB	06FB
C	060C		062C	063C	064C	065C	066C	067C	068C	069C	06AC	06BC	06CC	06DC	06EC	06FC
D	060D		062D	063D	064D	065D	066D	067D	068D	069D	06AD	06BD	06CD	06DD	06ED	06FD
E	060E	061E	062E	063E	064E	065E	066E	067E	068E	069E	06AE	06BE	06CE	06DE	06EE	06FE
F	060F	061F	062F	063F	064F		066F	067F	068F	069F	06AF	06BF	06CF	06DF	06EF	06FF

# *Variant Management System*



- 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**

**Requirements**

- 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

- 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

# Variant Management System

## Concepts (1): One key for all 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.

Label	Approximately # of variants
اتصال	300
اتصالات	6,000
الاتصالات	60,000
هيئة-الاتصالات	2,879,999
هيئة-الاتصالات-وتقنية-المعلومات	82,944,000,000

- 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).





# Variant Management System

## Concepts (2): Character Position



### ■ 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 considers character position when deciding if 2 code points are variants or not.

### – Example (هدهد):

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

هدهد	هدهد	هدهد	هدهد
هدهد	هدهد	هدهد	هدهد
هدهد	هدهد	هدهد	هدهد
هدهد	هدهد	هدهد	هدهد

### – Result:

- More Accuracy

# Variant Management System

## Concepts (3): Single Input Device



- 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 (كلى):
    - 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)

LANGUAGE	UNICODE	LABEL
N/A	(U+0643) (U+0644) (U+06CC)	كلى
N/A	(U+06A9) (U+0644) (U+0649)	كلى

LANGUAGE	UNICODE	LABEL
Arabic	(U+0643) (U+0644) (U+0649)	كلى
Urdu	(U+06A9) (U+0644) (U+06CC)	كلى

# Variant Management System

## Concepts (4): International Reachability



Visit our website:

کلی.موقع

قم بزيارة موقعنا:  
کلی.موقع



ک (0643)

ی (0649)

کلی.موقع



ک (06A9)

ی (06CC)

کلی.موقع

# *Variant Management System*

## *Concepts (4): International Reachability*



### ■ 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 “کلی” (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 “کلی” (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** pointy of view (based on input devices used by other language communities).





# Variant Management System

## Concepts (5): Simple User Interface



### ■ 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.



# *Variant Management System*

## *Requirements (1/2)*

- 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

# Variant Management System Requirements (2/2)



- 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

# SaudiNIC's Variants Management System

Unicode  
+  
IDNA  
+  
Supported  
Language(s)

Registry  
Language  
Tables

Variants  
Tables

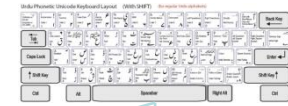
Master Key  
Algorithm

Domain name  
Master Key  
All Variants

Filters

- Requested Domain name
- Master Key
- Allocate-able Variants
  - ✓ Must be allocated Variants
  - ✓ Predicted Desired Allocate-able Variants
  - ✓ Predicted Undesired Allocate-able Variants
- Blocked Variants

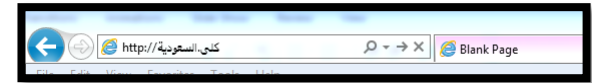
## Different Input Devices



Registrant



User



# SaudiNIC 's VMS

## Master Key



- 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)
    - كلي (U+0643) (U+0644) (U+064A)
    - كلي (U+06A9) (U+0644) (U+064A)
    - كلى (U+06A9) (U+0644) (U+06CD)
    - كلے (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 patters 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: (\*آل، آل\*, \*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.



# SaudiNIC's VMS

## Variant Filters



### ▪ Sample of our variant rules ( 21+ rules):

#### – AlefMadaEnd

- Input: خطأ-ظماً
- Filtered: خطأ-ظماً, خطأ-ظماً, خطأ-ظماً ..etc

#### – AlefHamzaDownEnd

- Input: خطأ-ظماً
- Filtered: خطأ-ظماً, خطأ-ظماً, خطأ-ظماً ..etc

#### – Alf-Altareef:

- Input: القرآن
- Filtered: القرآن, القرآن, القرآن

#### – Alef-letter-Alef

- Input: رايات
- Filtered: رايات, رايات, رايات

#### – .. etc.

**Note: Filtered variants are still can be allocated**

# SaudiNIC's VMS

## Variant Filters



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

label:

LABEL
الخطا-الظما
الخطا-الظما
الخطا-الظما
الخطا-الظما

### Results:

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

### Statistics Summary:

Total Variants	9999
I. Must be Allocated Variants (International Reachability )	0
II. Desired Variants	3
III. Not desired Variants	252
IV. Blocked Variants	9744

### I. Input:

LANGUAGE	UNICODE	LABEL
Arabic, Persian, Urdu, Malay, Pashto	(U+0627) (U+0644) (U+062E) (U+0637) (U+0627) (U+002D) (U+0627) (U+0644) (U+0638) (U+0645) (U+0627)	الخطا-الظما

# SaudiNIC's VMS

## Variant Filters



### Easy interface for registrants:

معلومات عن النطاق

اسم النطاق أمانة-مكة-المكرمة .السعودية

مثال للشبهات

اسم النطاق باستخدام أمانة مكة المكرمة .السعودية

المسافات (بدون): أدخل اسم النطاق مستخدماً "الفراغ" للفصل بين الكلمات (من دون شروط) شروط

المثال الأول: أمانة-مكة-المكرمة  
المثال الثاني: امانة-مكة-المكرمه  
المثال الثالث: امانه-مكه-المكرمه  
المثال الرابع: امانه-مكة-المكرمة

عرض الأمثلة

قائمة الشبهات الجديدة

حذف جميع الشبهات

الشبيه الأول .السعودية

الشبيه الثاني .السعودية

الشبيه الثالث .السعودية

# 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)
  - <http://nic.sa/en/view/doc64>
  - [http://arabic-domains.org/adn\\_tools/mk/index.php](http://arabic-domains.org/adn_tools/mk/index.php) (Demo)
  - Note: Will be updated to reflect new enhancements
  
- **Best practices** for managing Arabic domain names registries
  - Available soon

*Thank you*

شكراً

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



سجل.السعودية

nic.sa

هيئة الاتصالات وتقنية المعلومات  
Communications and Information Technology Commission



هيئة-الاتصالات.السعودية

citc.gov.sa