SaudiNIC:
Supporting Arabic Domain Names

Raed Alfayez, SaudiNIC
ICANN60, Abu Dhabi, Oct 2017
Agenda

➢ About SaudiNIC
➢ Introduction
➢ SaudiNIC’s major efforts
➢ What is missing?
About SaudiNIC

• Administering the domain name space under:
  – (sa) since 1995
  – (السعودية) since 2010.

• Operated by a government organization:
  – CITC (Communication and Information Technology Commission)

• Coordinating with regional and international bodies in order to present the local community needs.

• Leading the local and regional communities efforts towards supporting Arabic language in Domain Names since 2001 (more than 15 years of experience)
About SaudiNIC

50,813 Domain names
2LD/3LD Domain Names Distribution %

- com.sa: 55%
- sa: 25%
- pub.sa: 0.3%
- med.sa: 0.77%
- sch.sa: 1.4%
- edu.sa: 2.6%
- net.sa: 3%
- org.sa: 3.4%
- gov.sa: 3.6%
- 4.7%
Introduction: Arabic Language

• Ranked as the 5\textsuperscript{nd} language by native speakers in the world.
  – Native speakers: 295 million

• Considered as Official/Co-official language in 25 country

Source: http://en.wikipedia.org/wiki/Arabic_script
Introduction: Variants within the language
Introduction: Arabic Script

• The 2\textsuperscript{nd} 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
Arabic Script IDNs Major Issues

1. Combining Marks
2. Diacritics
3. World/label separators (space, ZWNJ, ZWJ, hyphen)
4. Digits
5. Confusing similar characters (e.g. variant tables)
6. Bidirectional
Main issues: Confusing Similar Characters

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

- 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 (کلى)!
  - Arabic (إنترنت) vs Arabic (انترنت)
SaudiNIC’s Major Efforts

Arabic IDN pilot projects
- Arab League (2005-2009)
- Language & Variant Tables

Tools, algorithms and solutions to manage variants:
- Master Key Algorithm
- Filters
- Variant Management System (VMS)

IDN Assessment Reports

Arabic Email Project (Raseel)
Arabic IDN pilot projects

• RFC: Linguistic Guidelines for the Use of the Arabic Language in Internet Domains
  – https://www.rfc-editor.org/rfc/rfc5564.txt

• For more information
Arabic IDN pilot projects

- Language & Variant Tables
SaudiNIC’s Major Efforts

Arabic IDN pilot projects
- Arab League (2005-2009)
- Language & Variant Tables

Tools, algorithms and solutions to manage variants:
- Master Key Algorithm
- Filters
- Variant Management System (VMS)

IDN Assessment Reports

Arabic Email Project (Raseel)
Tools and solutions: Compare Characters

- Display all code points of the whole Arabic script in one page
- Give the ability to compare code points based on their position
- It helped us to study the behavior of the code points and compare them against each other, in order to build our LT and VT.
Tools and solutions: **Master Key Algorithm**

- Secures the domain name space for the registry, speeds up lookup process and minimizes storage space:
  - Generates a unique key for a domain name label and all of its possible variants
  - The key can be used in the lookup process for both:
    - Domain name availability
    - Variants generation and allocation
- Supports multiple languages in a registry and it is easy to add a new language in the future
  - It requires a Language table (LT) and a Variant table (VT) for each supported language
- Provides automatic blocking of variants due to language mixing
- Supports defining variants based on character position
- Classify the relationship between variants (Exact /Typo/InterReach)
- ...etc

Check 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)
Tools and solutions: **Master Key Algorithm**

- **Exponential number of 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>
Tools and solutions: Filters (language based)

- **Goal:**
  - To reduce the huge size of allocate-able variants by intelligently identify and displaying 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 the word “cars”: “c”, “ca”, “ar”, “rs”, “s”
      - 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. (KACST Arabic Corpus)
    - We studied high-frequency patterns and then built some rules/filters based on them: (الـ، الـ*، الـ*، الـ*،... etc.)
      - We developed later a ranking system to arrange allocate-able variants based on weight given by each rule.
      - We have confirmed our findings with linguists and researchers.
Tools and solutions: Filters (language based)

- Sample of our variant rules (21+ rules):
  - AlefMadaEnd
    - Input: خطأ-ظماً
    - Filtered out: خطأ-ظماً, خطأ-ظماً, خطأ-ظماً etc
  - AlefHamzaDownEnd
    - Input: خطأ-ظماً
    - Filtered out: خطأ-ظماً, خطأ-ظماً, خطأ-ظماً etc
  - Alf-Altareef:
    - Input: القرآن
    - Filtered out: القرآن, القرآن, القرآن
  - Alef-letter-Alef
    - Input: رأيآت
    - Filtered out: رأيآت, رأيآت, رأيآت
    - .. etc.

Note
Filtered out variants are still can be allocated manually after some verification
Results:

**Master Key:** G43B G41M G18F G14I G42B G43M G41M G26F G43B G18F

**Statistics Summary:**

<table>
<thead>
<tr>
<th>Category</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Variants</td>
<td>3239</td>
</tr>
<tr>
<td>I. Must be Allocated Variants (International Reachability)</td>
<td>2</td>
</tr>
<tr>
<td>II. Desired Variants</td>
<td>4</td>
</tr>
<tr>
<td>III. Not desired Variants</td>
<td>28</td>
</tr>
<tr>
<td>IV. Blocked Variants</td>
<td>3205</td>
</tr>
</tbody>
</table>

I. Input:

<table>
<thead>
<tr>
<th>LANGUAGE</th>
<th>UNICODE</th>
<th>LABEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>(U+0645) (U+0643) (U+0629) (U+002D) (U+0627) (U+0644) (U+0645) (U+0643) (U+0631) (U+0645) (U+0629)</td>
<td>مكة المكرمة</td>
</tr>
</tbody>
</table>
## II. Must be Allocated Variants (2):

<table>
<thead>
<tr>
<th>LANGUAGE</th>
<th>UNICODE</th>
<th>TYPO/EXACT</th>
<th>LABEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persian, Malay, Pashto</td>
<td>(U+0645) (U+06A9) (U+0629) (U+002D) (U+0627) (U+0644) (U+0645) (U+06A9) (U+0631) (U+0645) (U+0629)</td>
<td>E</td>
<td>مكة-المكرمة</td>
</tr>
<tr>
<td>Urdu</td>
<td>(U+0645) (U+06A9) (U+06C3) (U+002D) (U+0627) (U+0644) (U+0645) (U+06A9) (U+0631) (U+0645) (U+06C3)</td>
<td>T</td>
<td>مكة-المكرمة</td>
</tr>
</tbody>
</table>
### III. Desired Variants (4):

<table>
<thead>
<tr>
<th>RULES</th>
<th>WEIGHT</th>
<th>LANGUAGE</th>
<th>UNICODE</th>
<th>TYPO/EXACT</th>
<th>LABEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>AllocatableInInputLanguage, SameLanguage, AlAtarif, Amlat, Sync, RuleOne, RuleTwo, RuleThree, RuleFour, RuleFive, RuleSix, RuleSeven, RuleEight, RuleNine, RuleTen</td>
<td>4</td>
<td>Arabic</td>
<td>ꞈ (U+0645) ꞈ (U+0643) ꞈ (U+0647) ꞈ (U+002D) ꞈ (U+0627) ꞈ (U+0644) ꞈ (U+0645) ꞈ (U+0643) ꞈ (U+0631) ꞈ (U+0645) ꞈ (U+0647)</td>
<td>ꞈ</td>
<td></td>
</tr>
<tr>
<td>AllocatableInInputLanguage, SameLanguage, AlAtarif, Amlat, Sync, RuleOne, RuleTwo, RuleThree, RuleFour, RuleFive, RuleSix, RuleSeven, RuleEight, RuleNine, RuleTen</td>
<td>2</td>
<td>Persian, Malay, Pashto</td>
<td>ꞈ (U+0645) ꞈ (U+06A9) ꞈ (U+0647) ꞈ (U+002D) ꞈ (U+0627) ꞈ (U+0644) ꞈ (U+0645) ꞈ (U+06A9) ꞈ (U+0631) ꞈ (U+0645) ꞈ (U+0647)</td>
<td>ꞈ</td>
<td></td>
</tr>
<tr>
<td>AllocatableInInputLanguage, SameLanguage, AlAtarif, Amlat, Sync, RuleOne, RuleTwo, RuleThree, RuleFour, RuleFive, RuleSix, RuleSeven, RuleEight, RuleNine, RuleTen</td>
<td>0</td>
<td>Arabic</td>
<td>ꞈ (U+0645) ꞈ (U+0643) ꞈ (U+0629) ꞈ (U+002D) ꞈ (U+0627) ꞈ (U+0644) ꞈ (U+0645) ꞈ (U+0643) ꞈ (U+0631) ꞈ (U+0645) ꞈ (U+0647)</td>
<td>ꞈ</td>
<td></td>
</tr>
<tr>
<td>AllocatableInInputLanguage, SameLanguage, AlAtarif, Amlat, Sync, RuleOne, RuleTwo, RuleThree, RuleFour, RuleFive, RuleSix, RuleSeven, RuleEight, RuleNine, RuleTen</td>
<td>0</td>
<td>Arabic</td>
<td>ꞈ (U+0645) ꞈ (U+0643) ꞈ (U+0629) ꞈ (U+002D) ꞈ (U+0627) ꞈ (U+0644) ꞈ (U+0645) ꞈ (U+0643) ꞈ (U+0631) ꞈ (U+0645) ꞈ (U+0629)</td>
<td>ꞈ</td>
<td></td>
</tr>
<tr>
<td>RULES</td>
<td>WEIGHT</td>
<td>LANGUAGE</td>
<td>UNICODE</td>
<td>TYPO/EXACT</td>
<td>LABEL</td>
</tr>
<tr>
<td>--------------------------------------------</td>
<td>--------</td>
<td>----------</td>
<td>---------------------------</td>
<td>------------</td>
<td>-------</td>
</tr>
<tr>
<td>AllocatableInInputLanguage, SameLanguage, AlAttarif, Amlat, Sync, RuleOne, RuleTwo, RuleThree, RuleFour, RuleFive, RuleSix, RuleSeven, RuleEight, RuleNine, RuleTen</td>
<td>-98</td>
<td>Arabic</td>
<td>(U+0645) (U+0643) (U+0629) (U+002D) (U+0622) (U+0644) (U+0645) (U+0643) (U+0631) (U+0645) (U+0629)</td>
<td>T</td>
<td>مكة-المكرمة</td>
</tr>
<tr>
<td>AllocatableInInputLanguage, SameLanguage, AlAttarif, Amlat, Sync, RuleOne, RuleTwo, RuleThree, RuleFour, RuleFive, RuleSix, RuleSeven, RuleEight, RuleNine, RuleTen</td>
<td>-100</td>
<td>Arabic</td>
<td>(U+0645) (U+0643) (U+0647) (U+002D) (U+0622) (U+0644) (U+0645) (U+0643) (U+0631) (U+0645) (U+0629)</td>
<td>T</td>
<td>مكة-المكرمة</td>
</tr>
<tr>
<td>AllocatableInInputLanguage, SameLanguage, AlAttarif, Amlat, Sync, RuleOne, RuleTwo, RuleThree, RuleFour, RuleFive, RuleSix, RuleSeven, RuleEight, RuleNine, RuleTen</td>
<td>-98</td>
<td>Arabic</td>
<td>(U+0645) (U+0643) (U+0629) (U+002D) (U+0622) (U+0644) (U+0645) (U+0643) (U+0631) (U+0645) (U+0629)</td>
<td>T</td>
<td>مكة-المكرمة</td>
</tr>
<tr>
<td>AllocatableInInputLanguage, SameLanguage, AlAttarif, Amlat, Sync, RuleOne, RuleTwo, RuleThree, RuleFour, RuleFive, RuleSix, RuleSeven, RuleEight, RuleNine, RuleTen</td>
<td>-100</td>
<td>Arabic</td>
<td>(U+0645) (U+0643) (U+0647) (U+002D) (U+0622) (U+0644) (U+0645) (U+0643) (U+0631) (U+0645) (U+0629)</td>
<td>T</td>
<td>مكة-المكرمة</td>
</tr>
<tr>
<td>AllocatableInInputLanguage, SameLanguage, AlAttarif, Amlat, Sync, RuleOne, RuleTwo, RuleThree, RuleFour, RuleFive, RuleSix, RuleSeven, RuleEight, RuleNine, RuleTen</td>
<td>-98</td>
<td>Arabic</td>
<td>(U+0645) (U+0643) (U+0629) (U+002D) (U+0622) (U+0644) (U+0645) (U+0643) (U+0631) (U+0645) (U+0629)</td>
<td>T</td>
<td>مكة-المكرمة</td>
</tr>
<tr>
<td>AllocatableInInputLanguage, SameLanguage, AlAttarif, Amlat, Sync, RuleOne, RuleTwo, RuleThree, RuleFour, RuleFive, RuleSix, RuleSeven, RuleEight, RuleNine, RuleTen</td>
<td>-100</td>
<td>Arabic</td>
<td>(U+0645) (U+0643) (U+0647) (U+002D) (U+0622) (U+0644) (U+0645) (U+0643) (U+0631) (U+0645) (U+0629)</td>
<td>T</td>
<td>مكة-المكرمة</td>
</tr>
</tbody>
</table>
SaudiNIC’s VMS

• An easy and stable variant management system:
  • No language mixing (utilizing the powerful tools: Language tables)
    – control input via the user interface
    – help identifying “must-be-allocated” variants for reachability purposes.
    – tremendously reduce the number of unnecessary allocateable variants
    – protect the TLD-space.
  • Master Key algorithm
    – Easily manage the whole variants list with one unique identifier
    – Speed up the lookup process
    – Eliminate the need of saving all possible variants
  • Must be allocated variants
    – For reachability purposes, “must-be-allocated” variants should be
      generated and activated automatically by the registry, so that: registered
      domain name is accessed regardless of the input devices (language table)
      being used by the navigator users.
  • Filters
    – To identify desired allocatable variants
For reachability purposes, variants should be addressed to be activated automatically by the registry, so that:

- A registered domain name is accessed regardless of the input devices (language table) being used by the navigator users.
- For example:
  - A user registered the domain “مكة” (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 “مکہ” (all characters from the Urdu language)
  - If the “must-be-allocated” variants were not allocated, delegated and hosted then the domain name will not be reachable.

Hence, reachability issue (based on input devices used by other language communities) should be carefully considered when defining variants (by language communities).
SaudiNIC’s VMS: Registrant will use his/her keyboard

مكة

U+0645
U+0643
U+0629

مکه

U+0645
U+0643
U+0647

مکه

U+0645
U+06A9
U+0647

مک

U+0645
U+06A9
U+06C1
<table>
<thead>
<tr>
<th>IDN</th>
<th>Total Variants</th>
<th>Allocatable</th>
<th>Blocked</th>
<th>Blocked due to Language Mixing</th>
</tr>
</thead>
<tbody>
<tr>
<td>مكة-المكرمة</td>
<td>3239</td>
<td>34</td>
<td>3205</td>
<td>3181 (99.25%)</td>
</tr>
<tr>
<td>القرآن-الكريم</td>
<td>11999</td>
<td>111</td>
<td>11888</td>
<td>11836 (99.56%)</td>
</tr>
<tr>
<td>هيئة-الإعلام</td>
<td>47999</td>
<td>81</td>
<td>47918</td>
<td>47764 (99.68%)</td>
</tr>
<tr>
<td>كهف-الياسمين</td>
<td>28799</td>
<td>65</td>
<td>28734</td>
<td>28680 (99.81%)</td>
</tr>
<tr>
<td>كهف-اكيا</td>
<td>21599</td>
<td>47</td>
<td>21552</td>
<td>21534 (99.92%)</td>
</tr>
</tbody>
</table>
SaudiNIC’s VMS: Language LGR and Script LGR

Language LGR (XML)

Script LGR (XML)

Secure Registry Domain Space

Limit variants...
SaudiNIC’s VMS: Easy interface for registrants

<table>
<thead>
<tr>
<th>Variant Management Form</th>
<th>Request Review</th>
<th>Request Review</th>
</tr>
</thead>
</table>

### Domain information

<table>
<thead>
<tr>
<th>Domain Name</th>
<th>مكة.المملكة</th>
</tr>
</thead>
</table>

### Must be allocated

1. **Variant 1**
   - Must be allocated
   - Arabic
   - مكة

2. **Variant 2**
   - Must be allocated
   - Persian
   - مكة

3. **Variant 3**
   - Must be allocated
   - Pashto
   - مكة

### Desired Variants

1. Desired Variants 1
   - مكة

2. Desired Variants 2
   - مكة

### Other Variants

1. Other Variant 1
SaudiNIC’s Major Efforts

Arabic IDN pilot projects
- Arab League (2005 - 2009)
- Language & Variant Tables

Tools, algorithms and solutions to manage variants:
- Master Key Algorithm
- Filters
- Variant Management System (VMS)

IDN Assessment Reports

Arabic Email Project (Raseel)
# IDN Assessment Reports

Conducted and Published a number of IDN Assessment Reports:

<table>
<thead>
<tr>
<th>Year</th>
<th>Reports</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>• <strong>IDN Top Level Domain Evaluations and Testing Report</strong></td>
</tr>
<tr>
<td></td>
<td>• with the cooperation of the Arabic Domain Name Pilot Project Team.</td>
</tr>
<tr>
<td>2010</td>
<td>• Arabic IDN Test Results for Browsers</td>
</tr>
<tr>
<td></td>
<td>• Mozilla Firefox &amp; <strong>Microsoft IE</strong></td>
</tr>
<tr>
<td>2014</td>
<td>• <strong>IDN Assessment Report</strong></td>
</tr>
</tbody>
</table>
SaudiNIC’s Major Efforts

Arabic IDN pilot projects
- Arab League (2005 - 2009)
- Language & Variant Tables

Tools, algorithms and solutions to manage variants:
- Master Key Algorithm
- Filters
- Variant Management System (VMS)

IDN Assessment Reports

Arabic Email Project (Raseel)
Raseel: An Arabic Email System

• Phase I (2010~2013):
  – A pilot project to test Arabic email addresses
  – Built before the EAI RFCs
    • Using a hack: convert the user part of the email address to Punycode
    • Implemented plugins for Outlook and Roundcube to display the Arabic addresses correctly.
  – Work with existing Email Servers and old RFCs.
Raseel: An Arabic Email System

• Phase II (2016+):
  – Built based on the new EAI RFCs using standard EAI addresses
    • Postfix, Horde/Roundcube and Archiveopteryx
  – Still in a beta version and not open for public.
  – Successful test internally and with Gmail and MS Outlook.
  – No need for plugins.
Raseel: An Arabic Email System
Almost 5 years since the EAI RFCs were published and until now there are almost no support (or very limited) in:

- Email servers (SMTP, IMAP, POP),
- Email providers (Gmail, Hotmail, Yahoo)
- Emails clients (Webmail, Application)

Need to have a protection mechanism for the user part of the emails addresses (similar to IDN variants)

<table>
<thead>
<tr>
<th>Arabic</th>
<th>Farsi</th>
</tr>
</thead>
<tbody>
<tr>
<td>رسالة@رسيل.السعودية</td>
<td>رسالة@رسيل.السعودية</td>
</tr>
<tr>
<td>Farsi Yeh (U+06CC)</td>
<td>Arabic Yeh(U+064A)</td>
</tr>
</tbody>
</table>

Automatic tools to configure and manage variants (Domain, User Accounts).

Boosting the adoption of the new EAI RFC by ISP and service/hosting providers.
WHAT IS MISSING?
Variants enablement must be done in every level

- **Registry**
  - Register and enable variants: مكة مكة مكة

- **DNS Hosting**
  - Configure DNS & add need RR (e.g. NS & A & CNAME) for:
    - xn--ogb5cf
    - xn--ogb9c4p
    - xn--hhb4rwc

- **Email Services**
  - Configure Email account and email aliases:
    - رائد@مكة رائد@مكة رائد@مكة

- **Web Hosting**
  - Configure web-server and account and aliases:
    - `<VirtualHost 10.10.10.10>
      DocumentRoot /makkah
      ServerName xn--ogb5cf
      ServerAlias xn--ogb9c4p
      ServerAlias xn--hhb4rwc
    </VirtualHost>`
Gift

- Published “SaudiNIC’s Best Practices in Supporting and Managing Arabic Domain Names”
  - http://www.nic.sa/docs/SaudiNIC_ADNBP.pdf
Thank you

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

هيئة الاتصالات وتقنية المعلومات
Communications and Information Technology Commission

سجل.السعودية
nic.sa

هيئة الاتصالات.السعودية
citic.gov.sa