المركز السعودي لمعلومات الشبكة
Saudi Network Information Center

هيئة الاتصالات وتقنية المعلومات
Communications and Information Technology Commission

SaudiNIC:

Securing Arabic Domain Names

Raed Alfayez, SaudiNIC

APTLD74, Tashkent, Sep 2018

Agenda

- About SaudiNIC
- About Arabic Script/Language
- Securing Arabic domains
- What's Missing?

About SaudiNIC

- Administering the domain name space under:
  - (.sa) since 1995
  - (.السعودية) Experimental since 2004, official since 2010.
- Operated by a government organization:
  - CITC (Communication and Information Technology Commission)
- Leading the local and regional communities efforts towards supporting Arabic language in Domain Names since 2001 (more than 15 years of experience)
- Participated in several TF & WG related to Arabic domains regionally and globally
- Developed: Tools, Pilot Projects, Reports & RFC related to Arabic domains

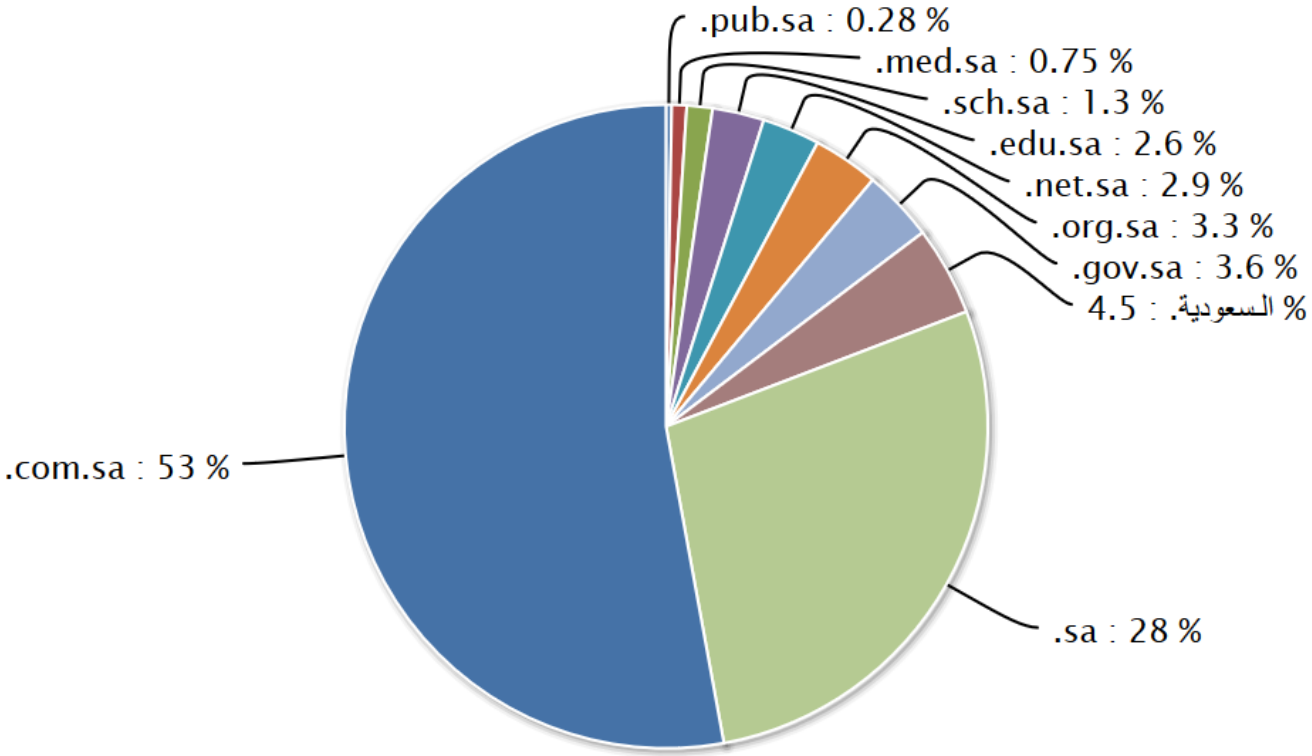
About SaudiNIC

55,850 Saudi domain names

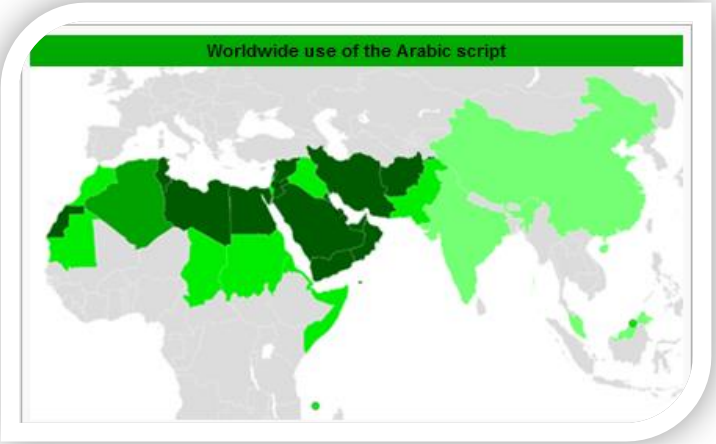
2,540 Arabic domains under (.السعودية)

466 Variant

2LD/3LD Domain Names Distribution %



Introduction: Arabic Script

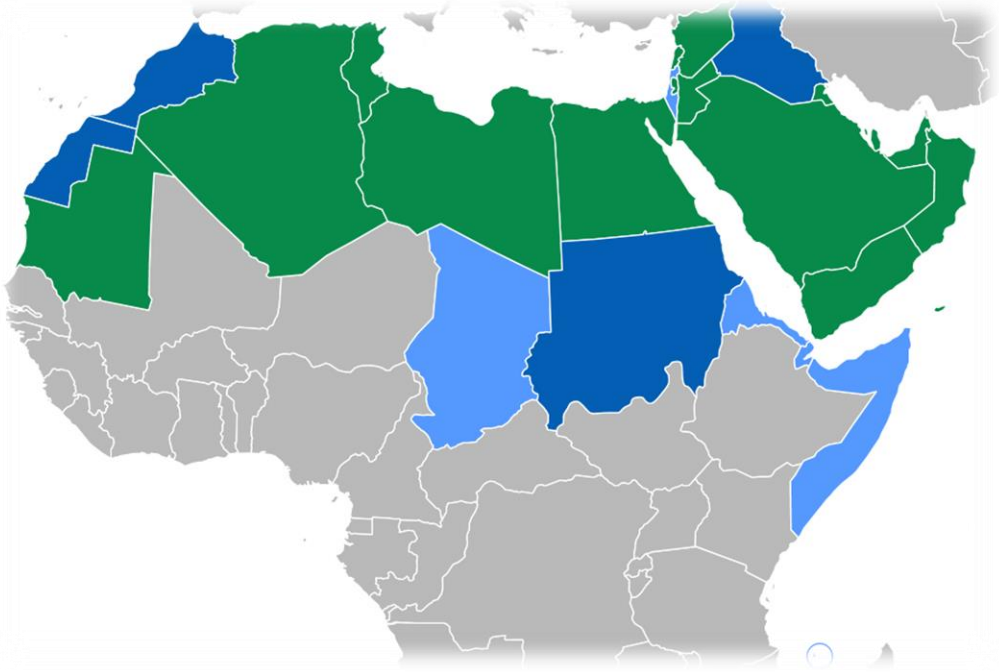


- The 2nd most widely used alphabetic writing system in the world
- Used by many languages such as:
  - Arabic, Urdu, Persian, Turkish, Kurdish, Pashto, …etc
- It is widely used by more than 43 countries
  - more than one billion potential users could be concerned in using Arabic script domain names.

Source: http://en.wikipedia.org/wiki/Arabic_script

Introduction: Arabic Language

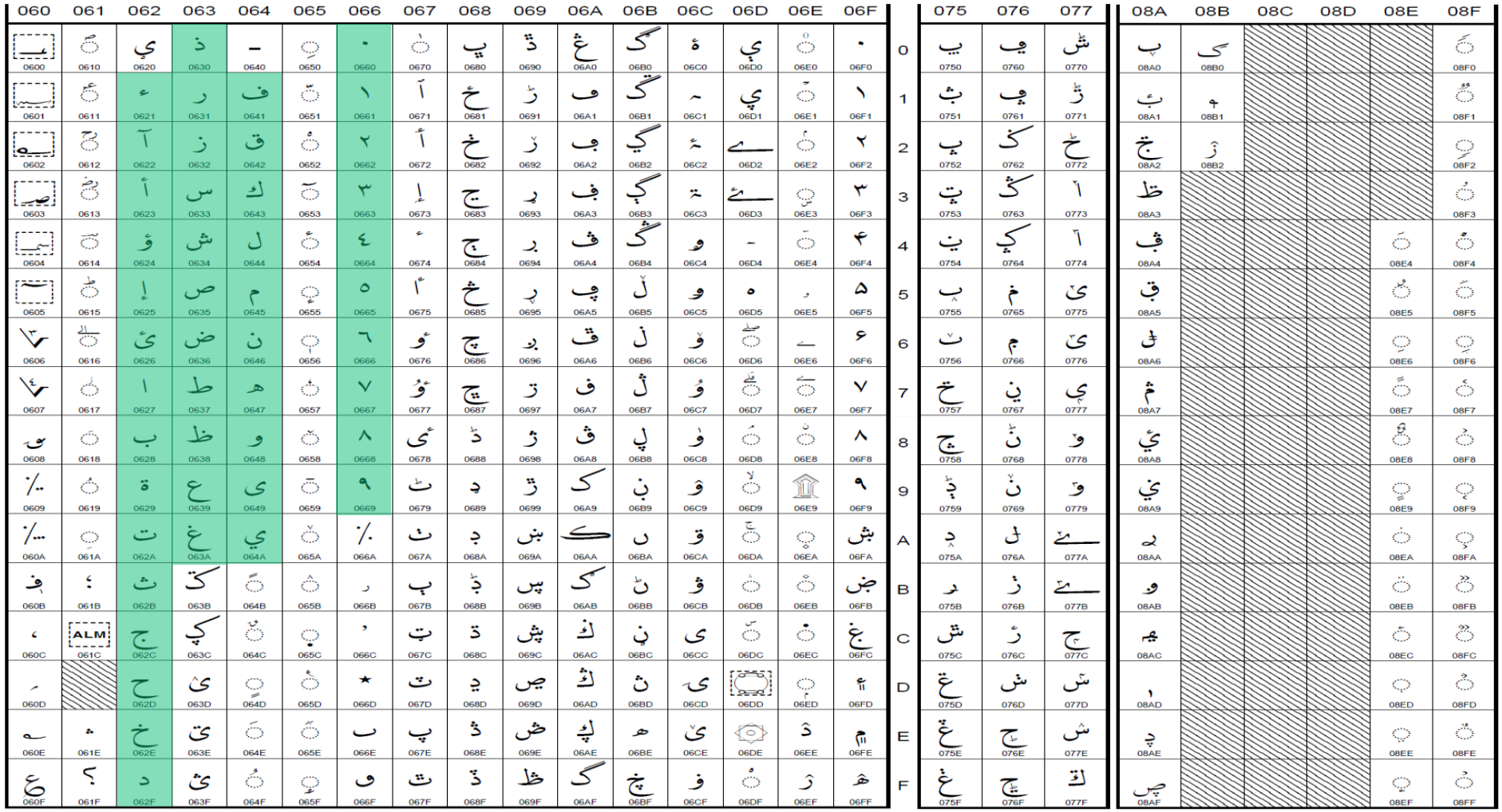
- Ranked as the 5th language by native speakers in the world.
- Considered as Official/Co-official language in 25 country, ~ 295 million native speakers.



Source: https://en.wikipedia.org/wiki/Arabic

Introduction: Arabic Language /Script

- The Arabic **language** is a subset of the Arabic **script**
  - **Script**: 0600..06FF, 0750..077F, 08A0..08FF **Language**: 0621..063A, 0641..064A, 0660..0669



Securing Arabic Domains

1

Identifying allowed code points

- Identifying allowed code points at language level and/or script level
- Not allowing unsecure/unsafe code points that may effect the stability of the registry domain space (e.g. control code points, combining marks, diacritics, Tatweel ..etc).

2

Identifying variants

- Variants within a language
- Variants across the whole script
- Reachability variants: So that a registered domain name can be accessed regardless of the input devices (language table) being used by the navigator users.

3

No script mixing

- A label can be constructed only by Arabic code points and digits

4

No language mixing

(utilizing the **powerful** tools: Language tables)

- control input via the user interface
- tremendously reduce the number of unnecessary/unrealistic variants
- Registry choice to protect the TLD-space.

5

No digits mixing

- Simplify the need for 3 digits sets in Arabic script
  - European digits
  - Arabic-Indic digits
  - Eastern Arabic-Indic digits

6

WLE Rules (language/scripts)

- Safeguard to handle some special cases (e.g. similarity in certain positions)

Securing Arabic Domains

(1) Identifying code points at both language and script level

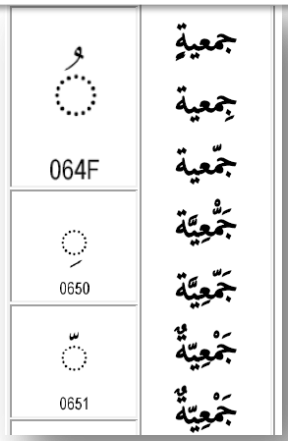
Include only:

✓ Allowed code points at language level and/or script level

Not included:

- Combining Marks
- Diacritics (Tashkeel)
- Special codes (ZWNJ, ZWJ)
- Unused/Historical

Non-spacing Marks



ZWNJ/ZWJ

Examples not using ZWNJ	Examples not using ZWNJ
طبل	طبل
input[0] = U+0637 input[1] = U+0628 input[2] = U+0644	input[0] = U+0637 input[1] = U+200c input[2] = U+0628 input[3] = U+0644

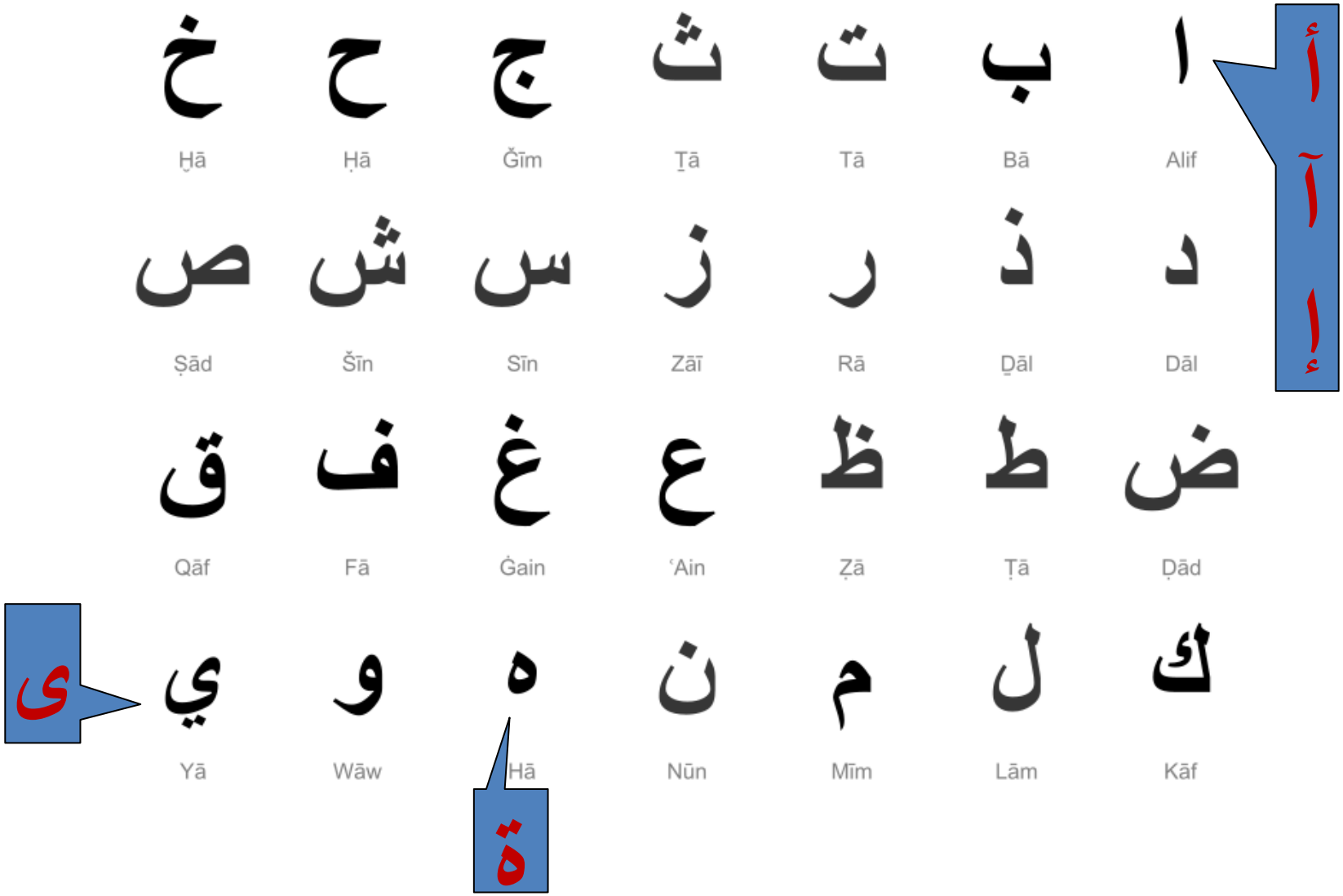
Combining Marks

ى	+	◌ٔ	=	ىٔ ئ ۓ	is confusing with	ئ ۓ ؤ
U+0649		U+0654		U+0649 U+0654		U+0626
Description: Alef Maksura + Hamza Above <> Yeh With Hamza Above						
Comments: This is a Unicode confusable!						

ى	+	◌ٔ	=	ىٔ ئ ۓ	is confusing with	ئ ۓ ؤ
U+06cc		U+0654		U+06cc U+0654		U+0626
Description: Farsi Yeh + Hamza Above <> Yeh With Hamza Above						
Comments: This is Unicode confusable!						

Securing Arabic Domains

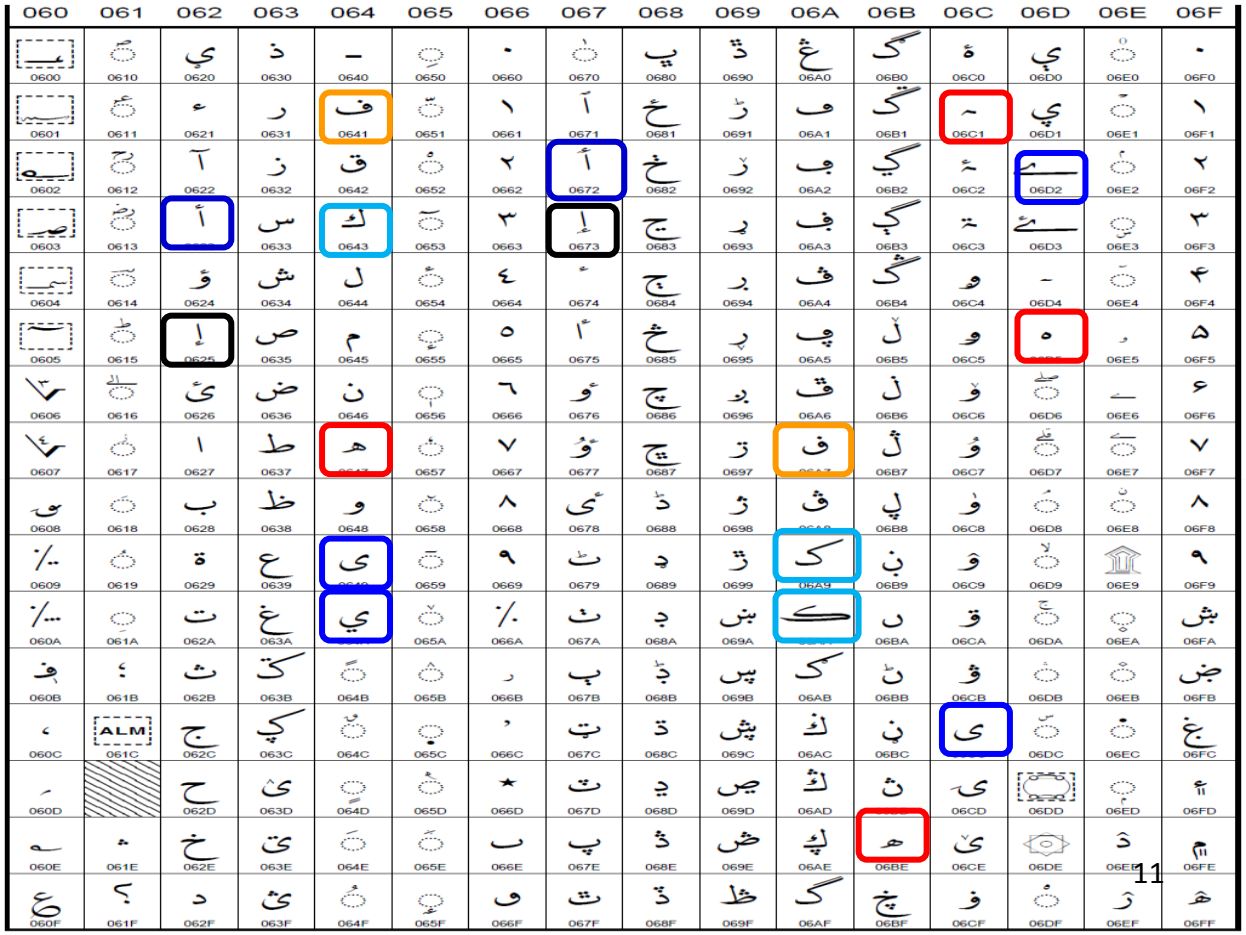
(2) Identifying variants (language level)



Securing Arabic Domains

(2) Identifying variants (script level)

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



Securing Arabic Domains

(2) Identifying variants (international reachability)

- For reachability purposes, variants should be addressed to be activated 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 "مکة" (all characters from the Urdu language)
    - If the "activated" variants were not allocated, delegated and hosted then the domain name will not be reachable!

LANGUAGE	UNICODE	LABEL
Arabic	(U+0645) (U+0643) (U+0629)	مكة
Persian, Malay, Pashto	(U+0645) (U+06A9) (U+0629)	مكة
Urdu	(U+0645) (U+06A9) (U+06C3)	مكة

Hence, reachability issue (based on input devices used by other language communities) should be carefully considered when defining variants (by language communities).

(0643) ك

(06A9) ک

Securing Arabic Domains
(3) No Script Mixing

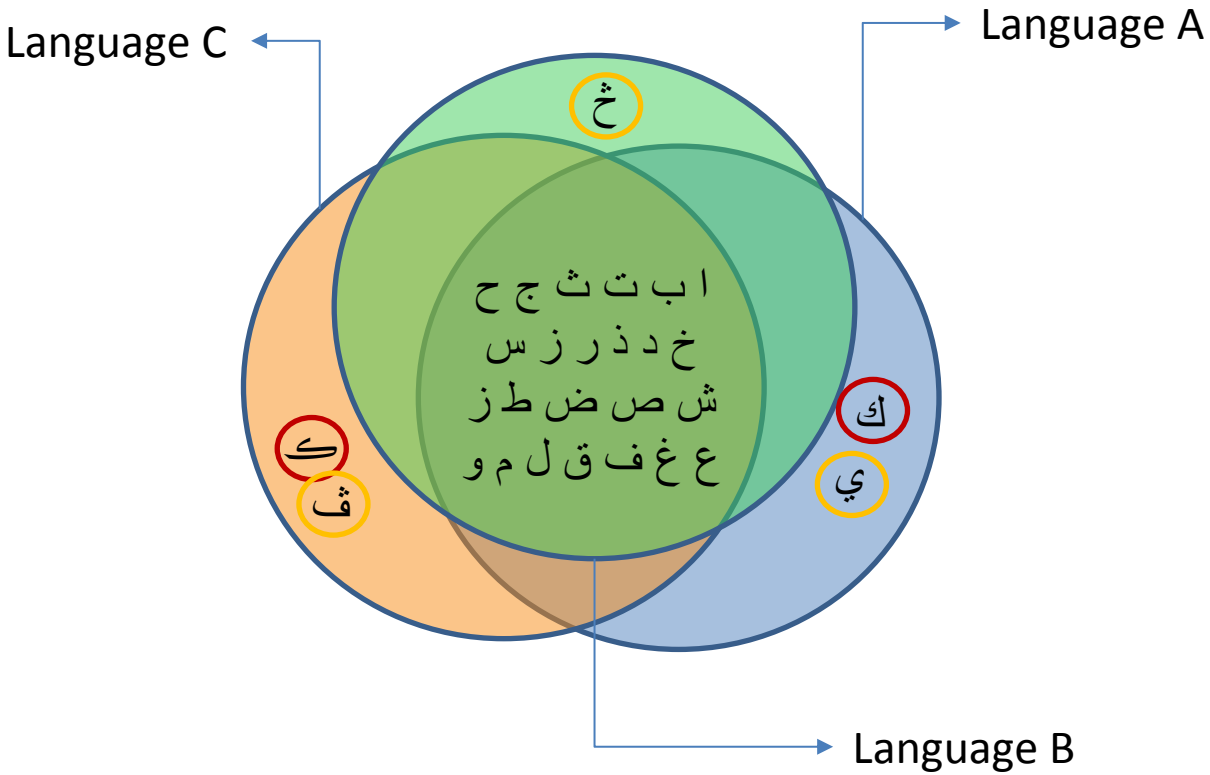
- Recommendation:
  - No mixing within a label or cross labels in the same domain

Format	Example
<FirstName><Arabic-LastName>	Salehالفلاني
<Arabic-LastName><FirstName>	الفلانيSaleh
<TLD>.<Arabic-Domain>	sa.رسيل
<Domain>.<Arabic-TLD>	raseel.السعودية

Securing Arabic Domains

(4) No language Mixing

- Not allowing code point that exists only in one language with another code points that exists only in another language in the same label
  - In the below example: ك & ک cant exist in the same label,
  - same thing for: ي& خ&ف



Securing Arabic Domains

(4) No language Mixing

IDN	Total Variants	Allocatable	Blocked	Blocked due to Language Mixing
مكة-المكرمة	3239	34	3205	3181 (99.25%)
القرآن-الكريم	11999	111	11888	11836 (99.56%)
هيئة-الإعلام	47999	81	47918	47764 (99.68%)
كهف-الياسمين	28799	65	28734	28680 (99.81%)
كهف-اكيا	21599	47	21552	21534 (99.92%)

Securing Arabic Domains

(5) No Digit Mixing

- Digits in the Arabic scripts:
  - European digits U+0030 .. U+0039 (0123456789)
  - Arabic-Indic digits U+0660 .. U0669 (٠١٢٣٤٥٦٧٨٩)
  - Eastern Arabic-Indic digits U+06F0 .. U+06F9 (۰۱۲۳۴۵۶۷۸۹)

#	Input string	Display
1	مؤتمر2009 conference2009	**Acceptable**: Pure European digits
2	مؤتمر٢٠٠٩ conference٢٠٠٩	**Acceptable**: Pure Arabic-Indic digits
3	مؤتمر۲۰۰۹ conference۲۰۰۹	**Acceptable**: Pure Eastern Arabic-Indic digits
4	مؤتمر٢٠٠9 conference٢٠٠9	**Not-Acceptable**: Mix between European digits & Arabic-Indic digits
5	مؤتمر2٠٠٩ conference2٠٠٩	**Not-Acceptable**: Mix between European digits & Arabic-Indic digits
6	مؤتمر٢٠٠٩ conference۲۰۰۹	**Not-Acceptable**: Mix between Arabic-Indic digits & Eastern Arabic-Indic digits
7	مؤتمر2٠٠٩ conference2٠٠٩	**Not-Acceptable**: Mix between European digits & Eastern Arabic-Indic digits

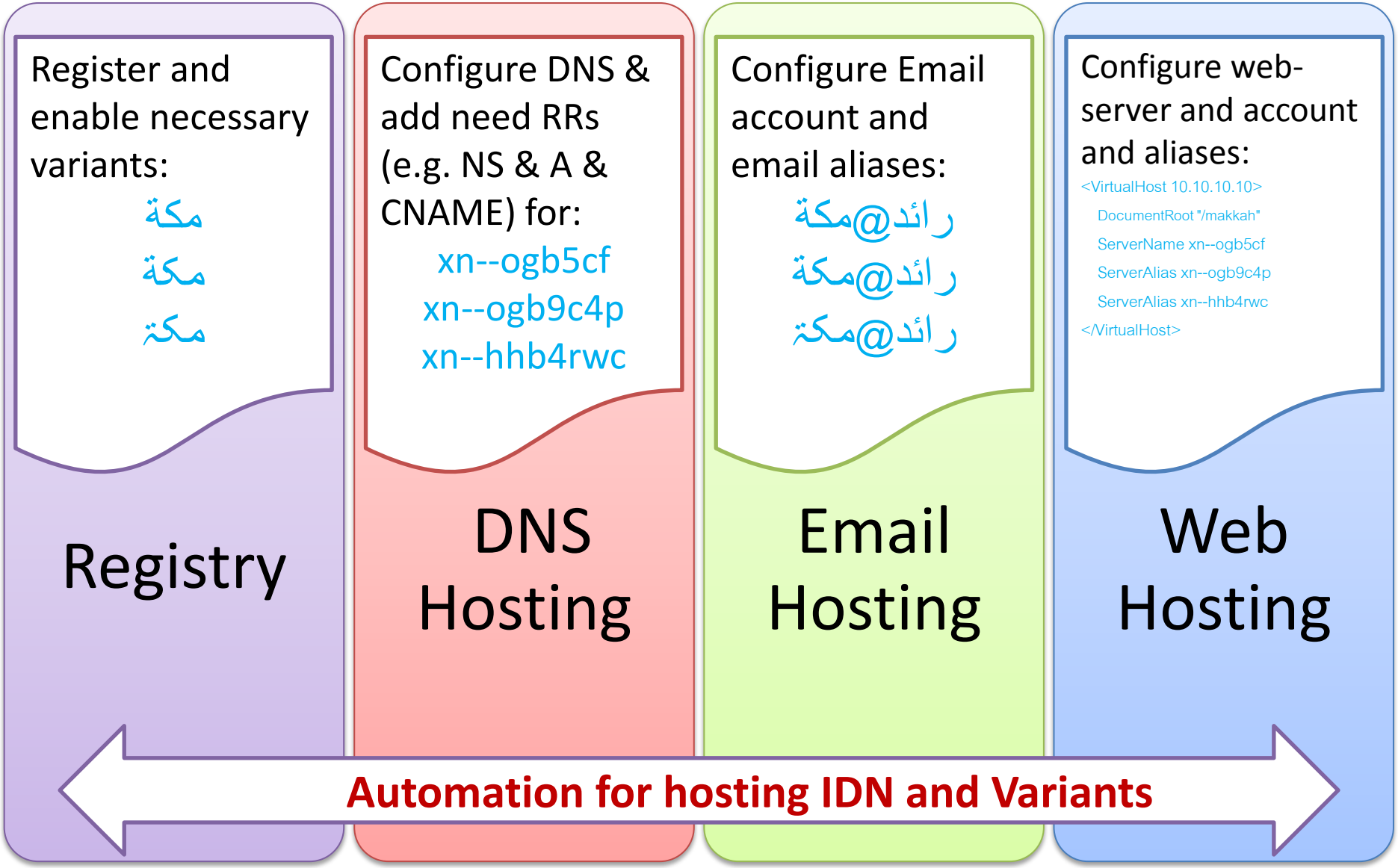
Securing Arabic Domains

(6) WLE Rules

- HEH and TEH MARBUTA
  - They are variants only in the final form (at the end)
  - جدة vs جده

- ALEF MAKSURA:
  - Preventing connected ALEF MAKSURA
  - سد vs ىىد

What's Missing?



Gift

- **SaudiNIC's Best Practices in Supporting and Managing Arabic Domain Names**

http://www.nic.sa/docs/SaudiNIC_ADNBP.pdf

Thank you

شكرًا

للمزيد من المعلومات يمكنكم زيارة:

For more information you can visit:

المركز السعودي لمعلومات الشبكة
Saudi Network Information Center

سجل.السعودية

nic.sa

هيئة الاتصالات وتقنية المعلومات
Communications and Information Technology Commission

هيئة-الاتصالات.السعودية

citc.gov.sa