



# Writing Domain Name Labels Using the Arabic Script

## ISSUES AND CONCERNS

Prepared by

**Dr. Abdulaziz H. Al-Zoman**  
**Director of SaudiNIC - CITC**  
**Chairman of Steering Committee**  
**Arabic Domain Name Pilot Project**

To be distributed during the  
First Global Workshop on Arabic Script Representation in IDNs  
March 30-31, 2008  
Grand Hyatt, Dubai, U.A.E.



## Table of Contents

Part I: Arabic Language.....	3
A. Linguistic Issues .....	4
1. Al-Tashkeel (Diacritics) .....	4
2. Kasheeda (Tatweel) .....	4
3. Character folding .....	4
4. Numbers .....	5
5. Connecting Multiple Words .....	5
6. Recommendations .....	6
B. Accepted Character Set Table .....	7
Part II: Arabic Script - Issues Need Further Investigations .....	9



# Part I

# Arabic Language



## A. Linguistic Issues

### 1. Al-Tashkeel (Diacritics)

Al-Tashkeel (diacritic) is a small sign that is usually put on top or under an Arabic letter for the purpose of correct pronunciation which may lead to a different meaning. Al-tashkeel is not a letter by itself but it is a mean to correctly pronounce a letter. It is not widely used except in case of the possibility of mispronouncing words that have the same letters but with different pronunciations, and hence having different meanings.

**Recommendation:** With respect to domain names, al-tashkeel can be supported only in the user interface but should not be stored in the zone file. Therefore, it can be stripped off at the preparation of internationalized strings ("stringprep") phase.

### 2. Kasheeda (Tatweel)

Kasheeda is not a letter. It is a horizontal line (like dash) used to lengthen the connection line between letters. It is used sometimes to enhance the display of Arabic words on screens or printouts.

**Recommendation:** Kasheeda should not be used in Arabic domain names.

### 3. Character folding

A character folding is the process where multiple letters (that may have some similarity with respect to their shapes) are folded into one shape. This includes:

- Folding Teh Marbuta and Heh at the end of a word.
- Folding different forms of Hamzah.
- Folding Alif Maksura and Yeh at the end of a word.
- Folding Waw with Hamzah and Waw.

Character folding is unacceptable in the Arabic language because it changes the meaning of the words and it is against the simplest spelling rules. Replacing a character with another character, which may have the same shape but different pronunciation, will give a different meaning. This will lead to have only one form (word) out many other forms of words that are made by all the combination of folded characters. Hence, the other forms will be masked by the common form.

Hand writing mixes between different characters (e.g., Heh and Teh-Marbuta) and this is due to laziness or weakness in spelling. However, this is not the case in published and printed materials. One of the motivations to support the Arabic language in domain names is to preserve the language particularly

with the spread of the globalization movement. Hence, character folding is working against this motivation since it is going to have a negative effect on the principles and ethics of the language.

**Recommendation:** Character folding should not be allowed.

#### 4. Numbers

In the Arab world, there are two sets of numerical digits used:

- Set I: (0, 1, 2, 3, 4, 5, 6, 7, 8, 9),  
Mostly used in the western part of the Arab world (al-maghrib al-arabi).
- Set II: (٠, ١, ٢, ٣, ٤, ٥, ٦, ٧, ٨, ٩),  
Mostly used in the eastern part of the Arab world (al-mashriq al-arabi).

There have been some suggestions to use Set I because it is thought that there is similarity (or confusion) between the Arabic zero (0) and the dot (.). But the differences appear clearly in publications. The zero is larger and is printed higher than the dot. Also, With respect to a domain name, it is quite easy to distinguish between the zero and the dot based on the context of the domain name. And since the two sets are used they should be both supported.

**Recommendation:** Both sets should be supported in the user interface and both are folded to one set (Set I) at the preparation of internationalized strings (e.g., "stringprep") phase.

#### 5. Connecting Multiple Words

In the Arab language words are separated by spaces. Connecting words without spaces is usually not acceptable. Therefore, a single space is the best word separator in an Arabic domain name with multiple words.

**Recommendation:** Space should be used to separate words if it is technically visible. Otherwise, it is recommended that multiple words are separated by the character "-" dash.

If the space is used as a word separator in Arabic domain names then it should be only a single space and it should not be used at the beginning or at the end of words.



## 6. Recommendations

Next Table lists Arabic Linguistic Committee recommendations regarding some linguistic issues.

<i>Issue</i>	<i>Recommendations</i>
<b>Tashkeel (Diacritics)</b>	<b>Tashkeel should not be allowed. However, if there is a need to allowed users to entered Tashkeel as part of a domain name then it should be stripped off by nameprep</b>
<b>Kasheeda</b>	<b>Kasheeda should be disallowed</b>
<b>Folding Teh Marbuta + Heh</b> <b>Folding different forms of Hamzah</b> <b>Folding Alif Maqsura+Ya</b>	<b>Folding should not be allowed</b>
<b>Numbers</b> <b>Arabic Zero</b>	<b>If it is technically possible, it is preferred to support both (Latin and Arabic) sets with folding to one set. Otherwise, Latin set is sufficient</b>
<b>Connecting Multiple Words</b> <b>Spaces</b>	<b>It is recommended that multiple words are separated by the character "-".</b>
<b>Mixing Latin and Arabic Characters</b>	<b>It is recommended that Arabic domain names be pure Arabic and they should not be mixed with other languages.</b>
<b>Special Characters (e.g., @, #, \$, %, ...)</b>	<b>It is recommended that Arabic domain names should follow the standard with respect to the use of special characters.</b>
<b>Accepted Character Set</b>	<b>It is recommended to use UNICODE 3.1. The following Unicode characters are accepted in Arabic domain names:</b> <b>U0621(hamza) until U063A (gheen)</b> <b>U0641 (feh) until U064A (yeh)</b> <b>(U0660 until U0669)٩-٠ Arabic numbers:</b> <b>Latin numbers: 0-9 (U0030 – U0039)</b> <b>Hyphen (U002D)</b> <b>Dot (U002E)</b> <b>Other than these characters are not allowed</b>



## B. Accepted Character Set Table

It is recommended to use only the following Unicode characters. The following codes are based on Unicode version 5.0.

### *Characters from Unicode Arabic Table (0600–06FF)*

0621	(ء)	ARABIC LETTER HAMZA
0622	(ا)	ARABIC LETTER ALEF WITH MADDA ABOVE
0623	(آ)	ARABIC LETTER ALEF WITH HAMZA ABOVE
0624	(ؤ)	ARABIC LETTER WAW WITH HAMZA ABOVE
0625	(إ)	ARABIC LETTER ALEF WITH HAMZA BELOW
0626	(ئ)	ARABIC LETTER YEH WITH HAMZA ABOVE
0627	(ا)	ARABIC LETTER ALEF
0628	(ب)	ARABIC LETTER BEH
0629	(ة)	ARABIC LETTER TEH MARBUTA
062A	(ت)	ARABIC LETTER TEH
062B	(ث)	ARABIC LETTER THEH
062C	(ج)	ARABIC LETTER JEEM
062D	(ح)	ARABIC LETTER HAH
062E	(خ)	ARABIC LETTER KHAH
062F	(د)	ARABIC LETTER DAL
0630	(ذ)	ARABIC LETTER THAL
0631	(ر)	ARABIC LETTER REH
0632	(ز)	ARABIC LETTER ZAIN
0633	(س)	ARABIC LETTER SEEN
0634	(ش)	ARABIC LETTER SHEEN
0635	(ص)	ARABIC LETTER SAD
0636	(ض)	ARABIC LETTER DAD
0637	(ط)	ARABIC LETTER TAH
0638	(ظ)	ARABIC LETTER ZAH
0639	(ع)	ARABIC LETTER AIN
063A	(غ)	ARABIC LETTER GHAIN
0641	(ف)	ARABIC LETTER FEH
0642	(ق)	ARABIC LETTER QAF
0643	(ك)	ARABIC LETTER KAF
0644	(ل)	ARABIC LETTER LAM
0645	(م)	ARABIC LETTER MEEM
0646	(ن)	ARABIC LETTER NOON
0647	(هـ)	ARABIC LETTER HEH
0648	(و)	ARABIC LETTER WAW
0649	(ى)	ARABIC LETTER ALEF MAKSURA



064A	(ي)	ARABIC LETTER YEH
0660	(٠)	ARABIC-INDIC DIGIT ZERO
0661	(١)	ARABIC-INDIC DIGIT ONE
0662	(٢)	ARABIC-INDIC DIGIT TWO
0663	(٣)	ARABIC-INDIC DIGIT THREE
0664	(٤)	ARABIC-INDIC DIGIT FOUR
0665	(٥)	ARABIC-INDIC DIGIT FIVE
0666	(٦)	ARABIC-INDIC DIGIT SIX
0667	(٧)	ARABIC-INDIC DIGIT SEVEN
0668	(٨)	ARABIC-INDIC DIGIT EIGHT
0669	(٩)	ARABIC-INDIC DIGIT NINE

***Characters from Unicode Basic Latin Table (0000–007F):***

0030	(0)	DIGIT ZERO
0031	(1)	DIGIT ONE
0032	(2)	DIGIT TWO
0033	(3)	DIGIT THREE
0034	(4)	DIGIT FOUR
0035	(5)	DIGIT FIVE
0036	(6)	DIGIT SIX
0037	(7)	DIGIT SEVEN
0038	(8)	DIGIT EIGHT
0039	(9)	DIGIT NINE
002D	(-)	HYPHEN-MINUS
002E	(.)	FULL STOP (Dot)



## Part II

# Arabic Script

## Issues Need Further Investigations