Arabic Domain Names

Dr. Abdulaziz H. Al-Zoman
Raed I. Al-Fayez
Arabic Domain Names Pilot Project
SaudiNIC, KACST, Saudi Arabia
20/2/2006

1 Introduction

Domain names are used widely by Internet users to locate resources on the Internet via a format that is easy to remember and understand. They are used instead of the numerical addresses which are known as Internet Protocol (IP) addresses. Hence, the main objective of using domain names is to ease and simplify the use of the Internet.

Despite the worldwide spread of the Internet, the Internet domain name system has not fully supported other languages to locate resources on the Internet. Users in non-English speaking countries, such as Arab users, are at a disadvantage. Using domain names in a language that is different from the users' native language defeats the main objective of having the domain name in characters rather than just numbers.

The Internet penetration in the Arab world is estimated to be 1.67 % and it was expected to be around 6.41 % by end of 2005, which is indeed very low. One of the obstacles facing the growth of Internet use in the Arab world is the language barrier. Many countries and nations are encouraging their people to use Internet, therefore it is important to ensure that the Internet supports the Arabic language, not only in web content but also in it's addresses.

2. Main Characteristics of Arabic Language

The Arabic language is the official language of 22 countries that are the members of the Arab League. Also, it is widely used by more than 43 Islamic countries. This means that there are more than one billion potential users who could be concerned in using Arabic domain names. The Arabic language has a number of characteristics, including the following:

a. It consists of 28 non-Latin-based alphabetic characters.
b. They are written from right to left. For example,
   Domain Names …. ... أسماء النطاقات
   □□□ writing direction writing direction □□ □.
c. The same characters could have several different shapes (e.g., يـ ، ـيـ ، ـي) depending on their position (beginning, middle, or end) within a word, and are probably conjugated with preceding and succeeding characters. These different shapes for a single character do not count as different code points but they are handled using different fonts.
d. Based on the above point "c", word separation – space between words – is inevitable to avoid joining of words in a way that would make them unreadable.
e. Tashkeel (diacritic) is a small sign that is usually put on top or under a character for the purpose of correct pronunciation which may result in a different meaning (e.g., Amman, Oman: عمّان ، عُمان). It is not a letter by itself but it is a mean to correctly pronounce a letter. Although not widely used, it is useful when applied to words that have the same letter combination. Tashkeel may change pronunciation and meaning.
f. Two numeral digits (0,1,2, ..9 and ٠،١،٢،...٩) currently used. They mean the same thing and written from left-to-right. But they have different code points.
g. Word abbreviations are not widely used and when used are in the form of characters separated by dots.

3. Why Arabic Domain Names are Needed?

It is required that the Arabic language is being used from the start of switching on the personal computer till getting information from the Internet. Thus, eliminating the need for the users to enter non-Arabic web (URL) addresses particularly if the sites are in Arabic. This led to the need of Arabizing domain names. There are a number of reasons why Arabizing is needed, such as:

Figure (1): Problems in correctly spelling domains

- Difficulty to reach Arabic sites using ASCII (English) domain names (pronunciation & spelling problems), see Figure (1) for example.
- There are a small percentage of Arabs who can read and write in English.
- There are many well-known Arabic names (in Arabic language) that need to be used in the Internet.
- Current ASCII-based domain names are incapable of representing (or substituting) Arabic characters.
- The current ICANN IDN solution (IDN.ASCII) is not suitable for languages that are not Latin-based (e.g., right-to-left or ideographic languages). It is not acceptable by the Arab users because of basically tow reasons: (1) it does not eliminate the language barrier problem (a user still needs to type in ASCII); (2) it introduces additional typing difficulties as it involves switching typing directions: right-to-left for Arabic, left-to-right for ASCII (English).
- As the trend nowadays is implementing and providing e-learning, e-government, and e-business, then it is important to provide the information and services using the user native language for all type of users: children, women, less-educated persons, senior citizens, … etc.
- Therefore, full Arabic domain names will encourage Arab users from all backgrounds to widely use the Internet

4. Discussions Points

- ICANN has issued two guidelines (Version 1.0 and 2.0) for implanting IDN. However, both guidelines are still based upon a handicapped IDN solution (i.e., IDN.ASCII) that does not support full IDN on a TLD level (i.e., IDN.IDN). These implementation scenarios are not suitable for languages that are not Latin-based, for example, languages written from right-to-left (e.g., Arabic, Farsi, Urdu, …) or ideographic languages (e.g., Chinese, Japanese, Korean, …).

- It seems that ICANN is focusing on IDN.ASCII solutions. This has introduced numerous problems due to the use of multiple languages under one label. A more practical approach, even if only for testing proposes, is to start the IDN support at a ccTLD level rather than at a gTLD. So that the TLD is written in a specific language (e.g. Arabic) that will be supported also on the SLD controlled by the same character set table. We call this Mono-IDN. A Mono-IDN represents an international domain name in which all labels in a domain name are expressed using the same language character set (e.g. "Arabic.Arabic" or "Chinese.Chinese", Urdo.Urdo", ...etc). While, Poly-IDN represents an international domain name in which each label in the domain name can be expressed using different language character set (e.g. "Arabic.English", "Chinese.English", "Urdo.Arabic" ...etc). Thus, we believe that while ICANN is focusing on supporting Poly-IDN and solving their problems, this should not prohibit the implementation of Mono-IDNs by ccTLD administrators who will develop their language guidelines and share them with ICANN.

- The Arabic script is the second most widely used alphabetic writing system in the world after the Latin alphabet. Originally developed for writing the Arabic language and carried across much of the Eastern Hemisphere. The Arabic script has been adapted to such diverse languages as Persian, Urdu, Turkish, Malay, Kurdish, and Swahili. Such adaptations may feature altered or new characters to represent phonemes that do not appear in Arabic phonology. A new character usually created based on modifying the basic shape of an existing Arabic character, for example, by adding more dots, and sometimes by adding the same character shape but with a different pronunciation or use. These additions have meaning to the new language but not to the original language. Appendix I (Column 1-3) lists the Arabic alphabet, Unicode code points, and character name. For each Arabic character, Column 4 lists some examples of the modified characters that are used by other Arabic script based languages which are easily confused with the original Arabic character. The following can be observed from Appendix I:

  a. Column 1 lists the accepted Arabic character set for an Arabic domain name.
  b. The general shape of a character is shown even if the character may have different shapes.
  c. Supplementary symbols, e.g., diacritics, numeric forms, punctuation, and special symbols, are not shown.
  d. Except for very few, most Arabic characters would easily be confused with some other characters from other languages (Persian, Urdu, Sindhi, Pashto, Kazakh, Kurdish , …)
  e. These confusing characters (i.e., have similar shapes) can not be folded together because they sound differently (i.e. have different phonemes) and are used differently.

  Therefore, it is impossible to have a single accepted character set for the Arabic script which is widely used by different languages. Thus, it is recommended that each Arabic script based language be treated separately, i.e., each language has it is own accepted character set that should be applied for domain name registration for that language only. The Arabic Domain Name Pilot Project is in the process of generating an Internet Draft regarding the support of Arabic language in domain names (http://www.arabic-domains.org/docs/guidelines.pdf).

  This suggestion (i.e., the use of Mono-IDN) is also true for other languages that may share the same letter shapes (e.g., Latin based languages). So having Mono-IDN will give any registry the control over the domain fishing problems and simplify its representation.

- The current punycode implementations have some problems when writing Arabic domain names that start or end with Arabic Indic digits, example, ١٢٣للتعليم or ٢٠٠٦أولومبيات. This problem does not exist when the digits are in the middle, e,g., الندوة٢٠٠٦العالمية

- Since word abbreviations are not widely used in Arabic then it is expected that we may have long Arabic domain

names. Based on the current experiments with the use of punycode we found the maximum length of Arabic domain name labels is around 33 characters (compare this with the 63 characters used in case of ASCII domains). This might introduce strong restriction on Arabic domain names, e.g.,: مؤسسة-فلان-الفلاني-للاستثمارات-العقارية-والتجارية

Appendix I

Arabic Char.	Uni-Code	Arabic Letter Name	Confusing Letters From Other Languages
ء	0621	Hamza	ٔ (0674), ﳓ (06FD)
آ	0622	Alef With Madda Above	ٱ (0671),
أ	0623	Alef With Hamza Above	ٱ (0671), أ (0672), ٵ (0675)
ؤ	0624	Waw With Hamza Above	ۥ (0676), ۄ (0677), ۆ (06C6), ۈ (06C8), ۉ (06C9), ۏ (06CF)
إ	0625	Alef With Hamza Below	إ (0673),
ئ	0626	Yeh With Hamza Above	ۓ (0678), ۓ (06CE), ۓ (06D3)
ا	0627	Alef	
ب	0628	Beh	ٻ (066E), ٹ (0679), پ (067E)
ة	0629	Teh Marbuta	ۃ (06C3), ۂ (06C0), ۂ (06C2)
ت	062A	Teh	ٺ (067A), ټ (067C)
ث	062B	Theh	ٽ (067D), ٹ (067F)
ج	062C	Jeem	ڃ (0683), ڄ (0684)
ح	062D	Hah	
خ	062E	Khah	ڂ (0681), څ (0682)
د	062F	Dal	ڈ (0688), ډ (0689), ډ (068A), ډ (068D)
ذ	0630	Thal	ڈ (0688), ڈ (068C)
ر	0631	Reh	ڑ (0691), ړ (0693), ړ (0695)
ز	0632	Zain	ڑ (0691), ڑ (0692), ڑ (0697)
س	0633	Seen	
ش	0634	Sheen	ﺵ (06FA)
ص	0635	Sad	ﺹ (069D)
ض	0636	Dad	ﺽ (069E), ﺿ (06FB)
ط	0637	Tah	
ظ	0638	Zah	ﻇ (069F)
ع	0639	Ain	
غ	063A	Ghain	ڠ (06A0)
ف	0641	Feh	ﻑ (06A1), ﭢ (06A2), ﭣ (06A3), ﭤ (06A4)
ق	0642	Qaf	ﻕ (066F), ﻗ (06A8)
ك	0643	Kaf	ک (06A9), ﻙ (06AA), گ (06AB), ﻛ (06AC), ﮒ (06AF), ﮓ (06B2), ﮔ (06B3)
ل	0644	Lam	ڵ (06B5), ﻟ (06B6), ڶ (06B7)
م	0645	Meem	ﻡ (06FE)
ن	0646	Noon	ڼ (06B9), ں (06BA), ڼ (06BC)
ه	0647	Heh	ھ (06BE), ہ (06C1), ہ (06D5)
و	0648	Waw	ۄ (06C4), ۅ (06C5), ۆ (06C6), ۇ (06C7), ۈ (06C8),
ى	0649	Alef Maksura	ى (06CC), ۍ (06CD), ې (06CE)
ي	064A	Yeh	ﭒ (067B), ﭘ (067E), ﻱ (06D0), ﻳ (06D1), ے (06D2),