Independent Submission
Request for Comments: 5564
Category: Informational
ISSN: 2070-1721

A. El-Sherbiny
M. Farah
UN-ESCWA
I. Oueichek
Syrian Telecom Establishment
A. Al-Zoman
SaudiNIC, CITC

February 2010

Linguistic Guidelines for the Use of the Arabic Language in Internet Domains

Abstract

This document constitutes technical specifications for the use of Arabic in Internet domain names and provides linguistic guidelines for Arabic domain names. It addresses Arabic-specific linguistic issues pertaining to the use of Arabic language in domain names.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for informational purposes.

This is a contribution to the RFC Series, independently of any other RFC stream. The RFC Editor has chosen to publish this document at its discretion and makes no statement about its value for implementation or deployment. Documents approved for publication by the RFC Editor are not a candidate for any level of Internet Standard; see Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc5564.

Table of Contents

1. Introduction

The Internet Engineering Task Force (IETF) issued in March 2003 a set of RFCs for Internationalized Domain Names (IDN) ([1], [2], and [3]), which were planned to become the de facto standard for all languages. In 2007 and 2008, the following working drafts were released that propose revisions to the IDNA protocol:

- Internationalized Domain Names for Applications (IDNA): Background, Explanation, and Rationale [5]

- Internationalized Domain Names in Applications (IDNA): Protocol [6]
- An updated IDNA criterion for right-to-left scripts [7]
- The Unicode code points and IDNA [8]

These documents are known collectively as "IDNA2008".

This document constitutes a technical specification for the implementation of the IDN standards in the case of the Arabic language. It will allow the use of standard language tables to write domain names in Arabic characters. Therefore, it should be considered as a logical extension to the IDN standards. It thus presents guidelines for the proper use of Arabic characters with the IDN standards in an Arabic language context.

This document reflects the recommendations of the Arab Working Group on Arabic Domain Names (AWG-ADN), established by the League of Arab States (LAS), based on standardisation efforts of the United Nations Economic and Social Commission for Western Asia (UN-ESCWA) and on that group's document, "Guidelines for an Arabic Internet Domain Name" [9]. This document is also in full harmony with recent rigorous discussions that took place within the major language communities that use the Arabic script in their languages.

This document provides guidelines for the ways Arabic characters may be used for registering Internet domain names and how linguistic-specific issues should be handled. A few rules are recommended for application at the protocol level.

The key words "MUST", "REQUIRED", "SHOULD", "RECOMMENDED", and "MAY" in this document are to be interpreted as described in RFC 2119 [4].

Comments on this document are solicited and should be addressed to the working group's mailing list at ESCWA-ICTD@un.org and/or the author(s).

2. Arabic Language-Specific Issues

The main objective of the creation of Arabic domain names is to have a vehicle to increase Internet use amongst all strata of the Arabic-speaking communities.

Furthermore, a non-user-friendly domain name would further add to the ambiguity and the eccentricity of the Internet to the Arabic-speaking communities, thus contributing negatively to the spread of the

Internet and leading to further isolation of these communities at the global level.

Hence, there have been intensive efforts (especially those spearheaded by Dr. Al-Zoman and contributed to by UN-ESCWA and its Arabic Domain Names Task Force (ADN-TF)) to reach consensus on a multitude of linguistic issues with the following goals:

- To define the accepted Arabic character set to be used for writing domain names in Arabic, which is the subject of this document.

- To define the top-level domains of the Arabic domain name tree structure (i.e., Arabic gTLDs and ccTLDs). This goal will be handled in a separate document.

The first meeting of the AWG-ADN, held in Damascus from January-February 2005, gave special attention to the following:

- Simplification of the domain names, whenever possible, to facilitate the interaction of the Arabic user with the Internet.

- Adoption of solutions that do not lead to confusion either in reading or in writing, provided that this does not compromise the linguistic correctness of used words.

- Mixing Arabic and non-Arabic letters in the domain name label is not acceptable.

2.1. Linguistic Issues

There are a number of linguistic issues that have been proposed with respect to the use of the Arabic language in domain names. This section will highlight some of them. This section is based on the papers of Dr. Al-Zoman ([10] and [11]) and on the report of the first meeting of AWG-ADN [12]. For details, the reader is encouraged to review these references.

2.1.1. Diacritics (Tashkeel) and Shadda

Tashkeel and Shadda are accent marks placed above or below Arabic letters to produce proper pronunciation. They are thus used to differentiate different meanings for different words with the same base characters.

Neither Tashkeel nor Shadda are permitted in zone files when registering domain names in the Arabic language, although they are permitted in the current edition of IDNA2008. They can be supported

or ignored, if necessary, in the user interface with local mappings and can be stripped before IDNA processing.

The following are their Unicode presentations:

U+064B ARABIC FATHATAN
U+064C ARABIC DAMMATAN
U+064D ARABIC KASRATAN
U+064E ARABIC FATHA
U+064F ARABIC DAMMA
U+0650 ARABIC KASRA
U+0651 ARABIC SHADDA
U+0652 ARABIC SUKUN

2.1.2. Kasheeda or Tatweel (Horizontal Character Size Extension)

Kasheeda (U+0640 ARABIC TATWEEL) must not be used in Arabic domain names and should be disallowed for Arabic language domain names. The Kasheeda is not a letter and does not have an effect on pronunciation. It is used to extend the horizontal length or change the shape of the preceding letter for graphical representation purposes in Arabic writing. Accordingly, it has no value for the writing of domain names. The same applies to all languages using the Arabic script. The authors recommend that it should be disallowed at the protocol level.

2.1.3. Character Folding

Character folding is the process where multiple letters (that may have some similarity with respect to their shapes) are folded into one shape. Examples of such Arabic characters include:

- Folding Teh Marbuta (U+0629) and Heh (U+0647) at the end of a word
- Folding different forms of Hamzah (U+0622, U+0623, U+0625, U+0627)
- Folding Alef Maksura (U+0649) and Yeh (U+064A) at the end of a word
- Folding Waw with Hamzah Above (U+0624) and Waw (U+0648)

With respect to the Arabic language, character folding is not acceptable because it changes the meaning of words and is against the principle of spelling rules. Replacing a character valid for use in domain names with another character also valid for use in domain names, which may have a similar shape, will give a different meaning. This will lead to only one word representing several words consisting

of all the combinations of folded characters. Hence, the other words will be masked by a single word [10].

Mis-spelling or handwriting errors do occur, leading to mixing different characters despite the fact that this is not the case in published and printed materials. One of the motivations of this effort is to preserve the language, particularly with the spread of the globalization movement. Within this context, character folding is working against this motivation since it is going to have a negative effect on the principle and ethics of the language. Technology should work to preserve the language and not to destroy it. Thus, character folding should not be allowed. The case of digits is treated in a separate section below.

2.2. Supported Character Set

A domain name to be written in Arabic must be composed of a sequence of the following UNICODE characters and the FULL STOP (u+002E) to separate the labels. These are based on UNICODE version 5.0. The tables below are constructed using an inclusion-based approach. Thus, characters that are not part of these tables are prohibited.

Unicode	Character Name
0621	ARABIC LETTER HAMZA
0622	ARABIC LETTER ALEF WITH MADDA ABOVE
0623	ARABIC LETTER ALEF WITH HAMZA ABOVE
0624	ARABIC LETTER WAW WITH HAMZA ABOVE
0625	ARABIC LETTER ALEF WITH HAMZA BELOW
0626	ARABIC LETTER YEH WITH HAMZA ABOVE
0627	ARABIC LETTER ALEF
0628	ARABIC LETTER BEH
0629	ARABIC LETTER TEH MARBUTA
062A	ARABIC LETTER TEH
062B	ARABIC LETTER THEH
062C	ARABIC LETTER JEEM
062D	ARABIC LETTER HAH
062E	ARABIC LETTER KHAH
062F	ARABIC LETTER DAL
0630	ARABIC LETTER THAL
0631	ARABIC LETTER REH
0632	ARABIC LETTER ZAIN
0633	ARABIC LETTER SEEN
0634	ARABIC LETTER SHEEN
0635	ARABIC LETTER SAD
0636	ARABIC LETTER DAD
0637	ARABIC LETTER TAH

0638	ARABIC LETTER ZAH
0639	ARABIC LETTER AIN
063A	ARABIC LETTER GHAIN
0641	ARABIC LETTER FEH
0642	ARABIC LETTER QAF
0643	ARABIC LETTER KAF
0644	ARABIC LETTER LAM
0645	ARABIC LETTER MEEM
0646	ARABIC LETTER NOON
0647	ARABIC LETTER HEH
0648	ARABIC LETTER WAW
0649	ARABIC LETTER ALEF MAKSURA
064A	ARABIC LETTER YEH
0660	ARABIC-INDIC DIGIT ZERO
0661	ARABIC-INDIC DIGIT ONE
0662	ARABIC-INDIC DIGIT TWO
0663	ARABIC-INDIC DIGIT THREE
0664	ARABIC-INDIC DIGIT FOUR
0665	ARABIC-INDIC DIGIT FIVE
0666	ARABIC-INDIC DIGIT SIX
0667	ARABIC-INDIC DIGIT SEVEN
0668	ARABIC-INDIC DIGIT EIGHT
0669	ARABIC-INDIC DIGIT NINE

Source: Supporting the Arabic Language in Domain Names [10]
Table 1: CHARACTERS FROM UNICODE ARABIC TABLE (0600-06FF)

Unicode	Digit Name
0030	DIGIT ZERO
0031	DIGIT ONE
0032	DIGIT TWO
0033	DIGIT THREE
0034	DIGIT FOUR
0035	DIGIT FIVE
0036	DIGIT SIX
0037	DIGIT SEVEN
0038	DIGIT EIGHT
0039	DIGIT NINE
002D	HYPHEN-MINUS

Source: Supporting the Arabic Language in Domain Names [10]
Table 2: CHARACTERS FROM UNICODE BASIC LATIN TABLE (0000-007F)

2.3. Arabic Linguistic Issues Affected by Technical Constraints

In this section, technical aspects of some linguistic issues are discussed.

2.3.1. Numerals

In the Arab countries, 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.

- Set II: (u+0660, u+0661, u+0662, u+0663, u+0664, u+0665, u+0666, u+0667, u+0668, u+0669) mostly used in the eastern part of the Arab world.

Both sets may be supported in the user interface; however, the rule of numeral homogeneity must be observed. The rule specifies that digits from the Arabic-Indic set of numerals (u+0660 to u+0669) should not be allowed to mix with ASCII digits (u+0030 to u+0039) within the same Arabic domain name label. Thus, the appearance of a digit from one set prevents the use of any other digit from the other set.

2.3.2. The Space Character

The space character is strictly disallowed in domain names, as it is a control character. Instead, the hyphen (Al-sharta, i.e., u+02D) is proposed as a separator between Arabic words to avoid confusion that can take place if the words are typed without a separator.

It is acceptable to use the hyphen to separate between words within the same domain name label.

3. Summary and Conclusion

The proposed guidelines are in full accordance with the IETF IDN standards and take into account Arabic-language-specific issues within a compromise between grammatical rules of the Arabic language and ease of use of that language on the Internet.

In summary, the guidelines specify that, in Arabic domain names:

- Accent marks (Tashkeel and Shadda) are not permitted.

- Character folding is not permitted.

- If a numeral from the Arabic-Indic or ASCII digit sets appears in a label, numeral homogeneity is required.

- The hyphen must be used as a word separator instead of space.

4. Security Considerations

No particular security considerations could be identified regarding the use of Arabic characters in writing domain names. In particular, any potential visual confusion between different character strings is avoided using the guidelines proposed in this document.

5. Acknowledgments

ESCWA ICT Division provided support and funding for the development of this document with the objective of reaching a standard for comprehensive Arabic domain names. Thanks are due to SaudiNIC for its continuous efforts in supporting the development of Arabic domain names.

John Klensin and Harald Alvestrand reviewed the document and provided useful editorial and substantive support to enrich it.

6. References

6.1. Normative References

[1] Faltstrom, P., Hoffman, P., and A. Costello, "Internationalizing Domain Names in Applications (IDNA)", RFC 3490, March 2003.

[2] Hoffman, P. and M. Blanchet, "Nameprep: A Stringprep Profile for Internationalized Domain Names (IDN)", RFC 3491, March 2003.

[3] Costello, A., "Punycode: A Bootstring encoding of Unicode for Internationalized Domain Names in Applications (IDNA)", RFC 3492, March 2003.

[4] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

6.2. Informative References

[5] Klensin, J., "Internationalized Domain Names for Applications (IDNA): Definitions, Background and Rationale", Work in Progress, September 2008.

[6] Klensin, J., "Internationalized Domain Names in Applications (IDNA): Protocol", Work in Progress, September 2008.

[7] Alvestrand, H. and C. Karp, "An updated IDNA criterion for right-to-left scripts", Work in Progress, July 2008.

[8] Faltstrom, P., "The Unicode Codepoints and IDNA", Work in Progress, July 2008.

[9] United Nations Economic and Social Commission for Western Asia (UN-ESCWA), "Guidelines for an Arabic Domain Name System (ADNS)", Work in Progress, November 2007.

[10] Al-Zoman, A., "Supporting the Arabic Language in Domain Names", October 2003, <http://www.arabic-domains.org/docs/NIC-docs/SupportingArabicDomainNmaes.pdf>.

[11] Al-Zoman, A., "Arabic Top-Level Domains", Paper presented in Expert Group Meeting on Promotion of Digital Arabic Content, the United Nations, Economic and Social Commission for Western Asia, Beirut, June 2003.

[12] League of Arab States, "Report of the first meeting of AWG-ADN, Damascus", February 2005, <http://www.arabic-domains.org/ar/intrnational-entites.php>.

Authors' Addresses

Ayman El-Sherbiny
Information and Communication Technology Division ESCWA
UN-House
P.O. Box 11-8575
Beirut
Lebanon

EMail: El-sherbiny@un.org

Mansour Farah
Information and Communication Technology Division ESCWA
UN-House
P.O. Box 11-8575
Beirut
Lebanon

EMail: farah14@un.org

Ibaa Oueichek
Syrian Telecom Establishment
Damascus
Syria

EMail: oueichek@scs-net.org

Abdulaziz H. Al-Zoman, PhD
SaudiNIC, General Directorate of Internet Services
IT Sector, CITC
King Abdulaziz City for Science and Technology
PO Box 6086
Riyadh 11442
Saudi Arabia

EMail: azoman@citc.gov.sa