

Arabic Domain Names (ADN)
Recommendations by the Egyptian Workgroup
November 2004

1. Internationalized Domain Names: History and Background

For many years, Internet users were predominantly English speaking. Recently this seems to have changed. According to the Internet statistics published by Global Reach, only 35.2% of the global Internet users are English speaking. Chinese speaking users account for a 13.7%, Spanish for 9% and Japanese for 8.4% [[Global Reach September 2004](#)]. This fact has driven the development of online local content in native languages, and has emphasized the need for Internationalized Domain Names (IDN), where non-English Internet users may access the web or email in their native language.

On the international level, a number of IDN researches and test-beds were started in the Asia Pacific region as early as 1998. This was followed by the formation of non-for-profit organizations to discuss general and language-specific IDN issues, such as the Joint Engineering Team (JET), the Multilingual Internet Names Consortium ([MINC](#)), the Chinese Domain Name Consortium ([CDNC](#)), the Arabic Internet Names Consortium (AINC), and others. In 2000 the Internet Engineering Task Force ([IETF](#)) formed the [IDN working group](#) to investigate and specify the requirements for supporting internationalized domain names. The work of the group resulted in the IETF issuing three RFCs ([3490-3491-3492](#)) to identify IDN standards and proposed technical solutions. This was followed by the Internet Corporation for Assigned Names and Numbers ([ICANN](#)) publishing, in 2003, the “[Guidelines for the Implementation of Internationalized Domain Names](#)” which define the general standards for IDN registration policies and practices. Subsequently, the Internet Assigned Numbers Authority ([IANA](#)) created the “[IDN Language Table Registry](#)” allowing TLD operators to register “Language Tables”, which IANA then makes available online for public access.

2. Arabic Domain Names: Status Quo

Arabic Domain Names (ADN) have attracted a lot of interest from the different countries in the Arab world. Various organizations, ISPs and vendors have initiated efforts to study and try to implement ADNs. Some Arab countries have even initiated pilot projects and test-beds for ADN, such as the “[GCC Pilot Project Implementing Arabic Domain Names](#)”. All those efforts have resulted in a number of documents and drafts that try to define ADN standards. Of those we list:

- [AINC Linguistic Committee survey results](#)
- [SaudiNIC’s Contributions and Experiences in ADN](#)

- [Research by the Information Technology Committee of the Egyptian Organization for Standardization and Egyptian Society for Arabizing Science](#)
- [Internet draft by the Arabic Domain Name Task Force \(ADNTF\) under ESCWA auspices](#)
- [Internet draft by the Arabic Information Engineers Task Force \(AIETF\)](#)

It is worth noting that any new work on ADN should study the existing drafts before coming up with new recommendations. In this document we try to summarize and compare recommendations as listed in the above drafts, in addition to those listed in the Syrian Work Team document, mainly with respect to the linguistic aspects of ADN. At the end we provide recommendations by the Egyptian Workgroup.

3. Different Aspects of ADN

Prior to introducing a comprehensive ADN system, a number of aspects must be addressed and resolved, to the best of the technology and without affecting Internet stability or the current operation of the DNS system. The only way to ensure a smooth and successful deployment of ADN, that is not resented by the global Internet community, is to tackle those different aspects via a phased approach. ADN aspects are divided into 3 main categories: linguistic, technical and policy issues.

Linguistic aspects include Arabic language-specific characteristics, such as diacritics, kashida, character folding, as well as numerals and the words delimiter. In other words the linguistic part tries to define a unified language table to be adopted by the ADN system. Those issues were extensively discussed in the above listed documents, yet there are a number of disagreements as described in the next section.

Technical aspects mainly address the ADN system topology and structure. This includes how applications are to deal with an ADN and whether it should be handled at the server-side or at the client-side. As per the IDN standards set by the IETF RFCs, namely “Internationalizing Domain Names in Applications” (IDNA), it has been chosen to deal with IDN by updating user applications only without any changes to DNS servers, resolvers or protocol elements. Technical aspects related to ADN in specific, are addressed by some of the above listed documents, and should be further investigated, studied and tested during the coming phases.

Policy aspects include a wide range of issues. A non-comprehensive list of those may tackle issues such as Arabic ccTLD naming, Arabic gTLD tree structure, reserved Arabic names as related to norms or religious and cultural concerns. A more broader list of issues may also include copyright and dispute resolution matters, as well as organizational issues such as who should be in charge of the language table and whether this should be handled nationally or multilaterally. In this respect it is worth noting that the Arabic script is also used by at least another ten languages. For this reason, all policy issues need to be studied and addressed with this fact in mind and with careful attention, not to create conflicts with other languages especially regarding gTLDs.

Some policy aspects are discussed into details by the above listed ADN drafts. For example the debate of having the domain name completely in Arabic versus allowing Arabic.English domain names. Also the mechanism for choosing Arabic ccTLD names. In addition to the Arabic gTLD structure, and whether it should be an Arabic version of the current English one or a completely new version that addresses the specific nature of the Arabic language. We believe that those issues are of high importance, but are too complicated to be agreed upon during this phase. Trying to wait for an agreement on those issues might only delay the process, therefore we believe that this phase should rather focus on reaching an agreement on the linguistic issues, specifically on the character set to be adopted in the Arabic language table, in order to push ADN a step further.

4. ADN Linguistic Aspects: Towards Initial Agreement

The below table summarizes recommendations with respect to the linguistic aspects, as specified by the above listed ADN drafts, as well as those by the Syrian work team.

Issue	ADNTF / SaudiNIC	AIETF	Syrian Work Team
Diacritics	Supported only in user interface Not stored in DNS records	Supported only in user interface Initial Phase: Not stored in DNS records Next Phases: Stored in DNS records SHADDA (U+0651) is treated differently and requires some algorithm	Not Supported in user interface Not stored in DNS records
Kashida TATWEEL (U+0640)	Not supported	Not supported	Not supported
Character Folding	Not supported	Only YEH (U+064A) folded to ALEF MAKSURA (U+0649)	Needs further discussion

Numerals	ARABIC-INDIC DIGITS (U+0660 to U+0669) supported only in user interface Not stored in DNS records Folded to ASCII DIGITS (U+0030 to U+0039)	Only ARABIC-INDIC DIGITS (U+0660 to U+0669) supported in user interface and stored in DNS records	Needs further discussion
Word Separator	HYPHEN-MINUS (U+002D) SPACE (U+0020) preferred but not supported due to technical limitations	SPACE (U+0020)	HYPHEN-MINUS (U+002D) SPACE (U+0020) preferred but not supported due to technical limitations
Adopted Character Set	UNICODE 3.1 : U+0621 to U+063A U+0641 to U+064A U+0660 to U+0669 U+0030 to U+0039 U+002D U+002E	UNICODE 3.1 : U+0020 U+0621 to U+063A U+0641 to U+064A U+064E to U+0651 U+0660 to U+0670	UNICODE 3.1 : U+0621 to U+063A U+0641 to U+064A U+0660 to U+0669 U+0030 to U+0039 U+002D U+002E

The following recommendations by the Egyptian Workgroup, are based on the below guidelines:

- Correct Arabic language rules are not compromised
- IETF IDN standards are followed
- The ultimate goal of having an easy-to-use ADN system (from user, registrant and registry perspective) is maintained to the best possible
- Provisional cyber-squatting problems are minimized to the best possible

The below recommendations are to be used only for Arabic domain names and are not meant for free Arabic text.

A. Diacritics

Supporting diacritics or tashkeel may seem, at first glance, to serve the best-interest of the Arabic language, by providing a means for the right pronunciation of a domain name and for avoiding any conflict in its meaning. Yet a closer look reveals the fact that a native Arabic-speaking user does not actually need diacritics to understand Arabic words, which he usually figures out from the context, and hence does not use

diacritics in his day-to-day writing (not even in official one). Therefore it is reasonable to conclude that supporting diacritics may only provide yet another barrier to the end-user and thus act as lost opportunities for registrants, who would lose most of their customers' hits due to probable misspelling of domain names. This would force registrants to reserve all different combinations of their registered names to avoid this lost traffic and to guard their domain names from cyber-squatters. Therefore, not supporting diacritics seems like a better solution. The limitation of this, is that words which share the same spelling, but have different meanings with different diacritics, will be registered as a domain name only once on a first come first serve basis. In addition the absence of context and diacritics in domains, might cause ambiguity in some cases. **We recommend ADN not to support diacritics, while adding the specific diacritics, which may change a word meaning: FATHA (U+064E), DAMMA (U+064F) and KASRA (U+0650) to the supported character set, and reserving their use during the current stage.** Future versions of ADN standards might choose to introduce their usage in specific cases.

B. Shadda

Linguistically SHADDA (U+0651) is considered a letter and not a diacritic, since it is a short form used for doubling letters. Yet we believe that the recommendation of supporting SHADDA with ADN, faces the same arguments that are against using diacritics. This is especially true, since SHADDA, like diacritics, is not used in day-to-day writing. It is mostly omitted by native Arabs, as it is pronounced inherently based on the context. Hence SHADDA should be treated the same as diacritics, while bearing in mind that the same limitation applies as listed above. **Therefore we recommend ADN not to support SHADDA (U+0651), while adding it to the supported character set, and reserving its use during the current stage.** Future versions of ADN standards might choose to introduce their usage in specific cases.

C. Kashida

Kashida or TATWEEL (U+0640) is just a display character which is optionally placed anywhere in the word without changing its meaning. Therefore not supporting the kashida won't affect the language and will ease the guessing and even the recall of domain names. **Hence we recommend ADN not to support TATWEEL (U+0640).**

D. Folding

Due to bad writing habits, some characters are commonly misused in place of each other, such as using HEH (U+0647) instead of TEH MARBUTA (U+0629). Promoting those spelling mistakes by folding such characters to each other for the sake of simplicity, is definitely unacceptable from a pure linguistic point of view, instead, every single Arabic letter needs to be supported and this should not be compromised.

The issue that has been repeatedly debated, is folding of YEH (U+064A), at word endings, to ALEF MAKSURA (U+0649). This debate is due to the fact that the form of the letter YEH at the end of the word is written in two different ways within the Arab world. In some countries it is written identical to ALEF MAKSOURA (where the difference is understood from the context), while in other countries it is differentiated

from ALEF MAKSOURA by adding two dots underneath. Regarding this debate, we referred to Arabic linguistic experts and to [the Academy of the Arabic Language in Cairo](#) (Magmaa Al Logha Al Arabia), which is the Egyptian authority in charge of the Arabic language. We were advised that YEH, in its form at the end of the word with dots below, is not an original Arabic letter, rather it was added to the language in some Arab countries, then it spread to other parts of the Arab world. The dots below were added to differentiate the YEH, at word endings, from the ALEF MAKSOURA, which was historically written with SUPERSCRIPT ALEF (U+0670) above. This is supported by the fact that old publications and manuscripts have no occurrence of the letter YEH (U+064A), written at the end of the word with dots below. **Therefore we recommend ADN not to support folding, except the folding of YEH (U+064A) at the end of the word to ALEF MAKSOURA (U+0649).**

E. Numerals

There are 2 sets of numbers that are used in the Arab world. The ARABIC-INDIC DIGITS (U+0660 to U+0669) and the ASCII DIGITS (U+0030 to U+0039). While some people debate that the ASCII set is the original set of numbers that was used by the Arabs, resolution of the Academy of the Arabic Language in Cairo, clearly states that the ARABIC-INDIC set is the original Arabic set of numbers, and that it should be used by all Arab countries. This resolution was also endorsed by the Union of Academies of the Arabic Language (Itihad Al Magame Al Arabia) and by the League of Arab States. **Therefore we recommend ADN to support only the ARABIC-INDIC set (U+0660 to U+0669).**

F. Word Separator

Arabic letters change their shape and form according to their position in the word and depending on the adjacent letters. Having no separator between words, cause the words to join in a way that makes them unreadable to the user. The existence of a word separator is inevitable and must be used between every two words, especially since acronyms are seldom in the Arabic language and when written are separated by dots. The most acceptable and appealing word separator is the SPACE (U+0020), just like in other languages. Yet using the SPACE with ADN is technically not possible, as it will cause the domain to fail, as specified by the IDNA standard.

Alternatively the HYPHEN-MINUS (U+002D) is suggested as a word separator as is the case with standard DNS. Yet in ASCII domain name labels, the separator is used occasionally, since multiple words in English can be easily joined or abbreviated. Hence using HYPHEN-MINUS is not compulsory with every English domain label consisting of multiple words. In Arabic, the HYPHEN-MINUS will be used, in each label that consists of more than one word, at least once or more, since by nature Arabic names consist of multiple words.

We strongly believe that a long term technical solution should accommodate for SPACE as a word separator, while conforming to IDN standards. **Therefore we recommend ADN to use the HYPHEN-MINUS (U+002D) as a word separator, at this stage, while working out a technical solution that would include the SPACE (U+0020).**

G. Arabic Language Table

The Arabic language table is an aggregation of the set of characters specified by the above recommendations. **Hence we recommend ADN to adopt the following Unicode 4.0 characters subset in the Arabic language table.**

U+002D	HYPHEN-MINUS
U+0621	ARABIC LETTER HAMZA
U+0622	ARABIC LETTER ALEF WITH MADDA ABOVE
U+0623	ARABIC LETTER ALEF WITH HAMZA ABOVE
U+0624	ARABIC LETTER WAW WITH HAMZA ABOVE
U+0625	ARABIC LETTER ALEF WITH HAMZA BELOW
U+0626	ARABIC LETTER YEH WITH HAMZA ABOVE
U+0627	ARABIC LETTER ALEF
U+0628	ARABIC LETTER BEH
U+0629	ARABIC LETTER TEH MARBUTA
U+062A	ARABIC LETTER TEH
U+062B	ARABIC LETTER THEH
U+062C	ARABIC LETTER JEEM
U+062D	ARABIC LETTER HAH
U+062E	ARABIC LETTER KHAH
U+062F	ARABIC LETTER DAL
U+0630	ARABIC LETTER THAL
U+0631	ARABIC LETTER REH
U+0632	ARABIC LETTER ZAIN
U+0633	ARABIC LETTER SEEN
U+0634	ARABIC LETTER SHEEN
U+0635	ARABIC LETTER SAD
U+0636	ARABIC LETTER DAD
U+0637	ARABIC LETTER TAH
U+0638	ARABIC LETTER ZAH
U+0639	ARABIC LETTER AIN
U+063A	ARABIC LETTER GHAIN
U+0641	ARABIC LETTER FEH
U+0642	ARABIC LETTER QAF
U+0643	ARABIC LETTER KAF
U+0644	ARABIC LETTER LAM
U+0645	ARABIC LETTER MEEM
U+0646	ARABIC LETTER NOON
U+0647	ARABIC LETTER HEH
U+0648	ARABIC LETTER WAW
U+0649	ARABIC LETTER ALEF MAKSURA
U+064A	ARABIC LETTER YEH
U+064E	ARABIC FATHA (reserved for future use)
U+064F	ARABIC DAMMA (reserved for future use)
U+0650	ARABIC KASRA (reserved for future use)
U+0651	ARABIC SHADDA (reserved for future use)
U+0660	ARABIC-INDIC DIGIT ZERO
U+0661	ARABIC-INDIC DIGIT ONE
U+0662	ARABIC-INDIC DIGIT TWO
U+0663	ARABIC-INDIC DIGIT THREE
U+0664	ARABIC-INDIC DIGIT FOUR
U+0665	ARABIC-INDIC DIGIT FIVE
U+0666	ARABIC-INDIC DIGIT SIX
U+0667	ARABIC-INDIC DIGIT SEVEN
U+0668	ARABIC-INDIC DIGIT EIGHT
U+0669	ARABIC-INDIC DIGIT NINE

* ARABIC LETTER YEH (U+064A) at the end of the word is folded to ARABIC LETTER ALEF MAKSURA (U+0649)

5. Egyptian Work Group Recommendations: Summary

Issue	Egyptian Work Group
Diacritics and SHADDA (U+0651)	Not supported FATHA (U+064E), DAMMA (U+064F), KASRA (U+0650) and SHADDA (U+0651) reserved for future use
Kashida or TATWEEL (U+0640)	Not supported
Character Folding	Not supported Only YEH (U+064A) at the end of the word folded to ALEF MAKSURA (U+0649)
Numerals	Only ARABIC-INDIC DIGITS (U+0660 to U+0669) supported
Word Separator	HYPHEN-MINUS (U+002D) supported Future solutions to work-out limitations on SPACE (U+0020)
Adopted Character Set	<u>UNICODE 4.0</u> : U+002D U+0621 to U+063A U+0641 to U+064A U+064E to U+0651 (<i>reserved for future use</i>) U+0660 to U+0669

6. Conclusion

Finally it is of utmost importance that all Arab countries agree on one single language table, to be published and registered at the IANA “IDN Language Table Registry”, for reference by TLD registries interested in Arabic domain names. This prevents having other Arabic language tables published by non-native Arabs, who are not necessarily aware of the Arabic language specifications. We believe that reaching an agreement on the linguistic issues, will push ADN a step further, allowing Arab countries to start registering Arabic domain names under their ccTLDs. Moreover this phase should also focus on identifying available technical solutions or developing new ones, based on IDN standards, to start ADN deployment as early as possible. We also believe that it is the ultimate goal of ADN to have the domain name completely in Arabic, yet we think that for the benefit of moving on, ADN registrations should start under the current domain tree structure. The next phase should focus on policy issues, especially on Arabic ccTLDs and gTLDs, as well as on technical issues, such as including the SPACE as a word separator.