VISUALLY MISLEADING CHARACTERS IN THE ARABIC URL

Thomas Milo

SUMMARY

The purpose of this report is to make an inventory of vulnerabilities of encoded Arabic in the context of assigning names on the Internet. There is a historical introduction, a note about the structure of the script and a list of patterns of confusability.

SOME HISTORICAL BACKGROUND

Arabic disambiguation marks, the dots and vowels, are not part of the original orthography. The dotted marks did exist, but in early papyrus fragments and Qurʾān manuscripts these are used sparingly and inconsistently. Popular explanations that at some point in the spread of Islam they were introduced to facilitate non-Arab readers therefore do not stand up to scrutiny, while an academic analysis and explanation has not yet been produced.
Without the familiar disambiguation marks, the full semantic burden shifts to the bare script skeleton, the *rasm*, as is the case with the oldest manuscripts. The structure of the script is still essentially the same today.

A NOTE ABOUT THE STRUCTURE OF THE ARABIC SCRIPT

The basic structure consists of a system of dot-less skeleton letters that we call archigraphemes. Archigraphemes can be classified in three subgroups:

1. Full archigraphemes

Each archigrapheme represents a group of graphically similar letters:

A	B	G	D	R	S	C	T	E	F	H	W
ا	ٮ	ح	د	ر	س	ص	ط	ع	ڡ	ه	و

2. Position-dependent archigraphemes

This group consists of graphemes, that only occur in final position, in non-final position they are represented by the archigrapheme indicated between brackets in the table below:

[B] -N	[B] -Y	[F] -Q
ں	ى	ٯ

3. Pseudo-archigraphemes

This is a small group of graphemes for which no variants with diacritics exist in Arabic. Outside Arabic, however, these graphemes also take diacritics in all positions. Therefore these graphemes are archigraphemes in this context:

K	L	M
ك	ل	م

Here is the full archigraphemic character set and its transliteration.

A	B	G	D	R	S	C	T	E	F	[F] -Q	K	L	M	[B] -N	H	W	[B] -Y
ا	ٮ	ح	د	ر	س	ص	ط	ع	ڡ	ٯ	ك	ل	م	ں	ه	و	ى

THE PATTERNS OF ENCODING OF VISUALLY MISLEADING CHARACTERS

General observations

Encoding and rendering of URL's takes place in a context where:

1. The font size in URL command lines is too small to disambiguate the various points and other diacritics on which the identity of letters depends. This is aggravated by the design of the typefaces, and font layout technology that cannot prevent clashes between skeleton and diacritics. URL typography is of a vulnerable quality, particularly regarding the disambiguating diacritics.

2. In the Arabic block of the Unicode Standard, regional graphic preferences ended up as encoded letters, resulting in multiple instances of conceptually, and sometimes even visually identical letters with imperceptible and sometimes even invisible differences in encoding.

Example: The word MKBH (makkiyah) is the feminine adjective of the place name Mecca and can be encountered in the context of any of the Arabic-scripted orthographies. This example will be elaborated below.

Mechanisms of confusion

Optional use of dots

Throughout the history of Arabic script, final Yeh and Heh occur with dots (Unicode character 064A Yeh and 0629 Teh Marbuta) and without dots (Unicode character 06CC Yeh and 0647 Heh). This is a present day instance where the dots' historical independence from the underlying letters survives strongest: مكية and مكيه are interchangeable just as مكي and مکی are interchangeable.

Applying or omitting dots is frequent through the ages of written and printed text transmission. In addition to that there are regional preferences that last to this day. E.g., Egypt prefers dot-less final Yeh: مکی rather than مكي . This leads to the use of a third Unicode for Yeh, 0649 Arabic Letter Alef Maksura, which again for the unsuspecting public is undistinguishable from the other encoded characters that can be used to represent Yeh: مكى .

Optional rotation of dots

In Arabic proper, but also in Persian, Ottoman and Urdu, only the *number* of dots (0, 1, 2 or 3) counts, not their orientation. In the history and in the every day reality of Arabic script the orientation of the 2-dot character is irrelevant. Therefore مكية and مكية are interchangeable. This particular variation did not produce a new encodable character.

However, the free variant of Yeh evolved into an encodable character, e.g., for Uyghur. Given the fact that font designs are free to use dot rotation as an ornamental aspect of Arabic script, the letter that the unsuspecting non-Uyghur public can only read as Yeh, can in fact be encoded in three ways: مكية (064A), مکیة (06CC) or مكېة (06D0).

In addition to that, there is Unicode character 067B 'ARABIC LETTER BEEH', that shares, like Yeh, the archigrapheme B in non-final position. In final position the form of these characters is different, but when used in middle position, it produces another indistinguishable variant: مكٻة (067B).

Free and bound calligraphic variation

Final Heh/Teh Marbuta of MKBH *makkiyyä* "Meccan" is ascending in the *naskh* style that is preferred in the Arab world: مكية , while there is a *free* calligraphic descending variant: مكية.

However, in the *nastaliq* script, the preferred style of the Iranian and Indian worlds, the descending Heh is the only option, a *bound* variant: مکیہ .

In order to adapt to user expectation by "Iranianising" *naskh,* a special character was encoded as 06C3: مكيۃ. This is a problem that should have been handled by a font change that slipped into the Unicode standard as a separate character.

Meanwhile, ways have been found to create fonts in the *nastaliq* style, the style without the ascending Heh. The resulting confusion is that both مكية and مكيۃ show up in *nastaliq* as مکیہ and مکیۃ , because the difference between such Unicode characters cannot be expressed by fonts in this style.

Style-dependent variation

Medial Kaf of MKBH makkiyyä "Meccan" differs between Arabic, Persian and Urdu input. Yet in the perception of the reader there is no difference. Final form of Kaf in the naskh style is distinguished from Lam by a miniature Kaf: ـــل – ـــك, while in the nastaliq style it is in all positions, including final position, distinguished from Lam with a slanted stroke: ككك – کلک . In order to "Iranianise" naskh, a special character was encoded: ک. However, the nastaliq style doesn't have a Kaf without tail, so both ككك and ککک show up in nastaliq as ککک and ککک : in this type of Arabic script, the difference between such stylistic Unicode characters cannot be expressed in any contextual position.

In these cases Unicode is overruled by script styles, e.g., the contrast between the ascending and the descending variant forms of Heh is neutralised in a single descending shape. From a reader's perspective, these letters are perceptually identical. Here the gap between an engineering and literate reading leads to confusion.

The patterns of confusion

1. Pattern One – the multiple-choice characters

The variation between the regional Arabic script styles is almost comparable to the variation within the combined Greek, Cyrillic and Latin scripts. The latter group was encoded in three separate blocks, which inadvertently lead to encoding functionally and visually identical letters as separate characters in different sections of the Unicode Standard. As a result, e.g., the domain ending .com, can theoretically be encoded in six patterns that are indistinguishable for the reader:

C		O		M		total variants
c	0063	o	006F	m	006D	
с	0441	о	043E			
		ο	03BF			
2		3		1		6

For a casual reader, all these variants read as /.com/:

.com .cοm .cоm .сom .сοm .соm
.com .cοm .cоm .сom .сοm .соm

Arabic example: the word MKBH مكية (*makkiyah*). The amount of potential confusion can be calculated in this form.

M		K		B		H		total variants
م	0645	ك	0643	ي	064A	ة	0629	
		ک	06A9	ی	06CC	ۃ	06C3	
				ې	06D0	ە	06D5	
				ٻ	067B	ه	0647	
	1		2		4		4	32

This product of this is 32 graphemically identical words with slight allographic, i.e., semantically irrelevant, variation, but each one with a different encoding (colours corresponding to those in the table):

مكيه مكيە مكيۃ مكية مكیه مكیە مكیۃ مكیة مكٻه مكٻە مكٻۃ مكٻة كېه مكېە مكېۃ مكېة مکيه مکيە مکيۃ مکية مکیه مکیە مکیۃ مکیة مکېه مکېە مکېۃ مکېة مکٻه مکٻە مکٻۃ مکٻة

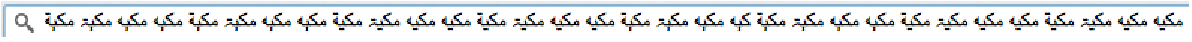
مكيه مكيە مكيۃ مكية مكیه مكیە مكیۃ مكیة مكٻه مكٻە مكٻۃ مكٻة كېه مكېە مكېۃ مكېة مکيه مکيە مکيۃ مکية مکیه مکیە مکیۃ مکیة مکېه مکېە مکېۃ مکېة مکٻه مکٻە مکٻۃ مکٻة

Graphemically identical means that for a casual reader, all these variants all read as /makkiyah/:

مكيه مكيە مكيۃ مكية مكیه مكیە مكیۃ مكیة مكٻه مكٻە مكٻۃ مكٻة كېه مكېە مكېۃ مكېة مکيه مکيە مکيۃ مکية مکیه مکیە مکیۃ مکیة مکېه مکېە مکېۃ مکېة مکٻه مکٻە مکٻۃ مکٻة

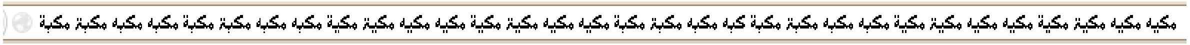
The following two examples show representative renderings for these characters as they the would occur on computer screens in the context where URL's are usually displayed. These fonts are designed not for clarity but to fit the vertical constraints of user interface elements, particularly at the expense of minute but semantically relevant graphic elements.

Command line on MacOSX



مكيه مكيە مكيۃ مكية مكیه مكیە مكیۃ مكیة مكٻه مكٻە مكٻۃ مكٻة كېه مكېە مكېۃ مكېة مکيه مکيە مکيۃ مکية مکیه مکیە مکیۃ مکیة مکېه مکېە مکېۃ مکېة مکٻه مکٻە مکٻۃ مکٻة

Command line on Windows

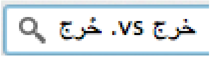


مكيه مكيە مكيۃ مكية مكیه مكیە مكیۃ مكیة مكٻه مكٻە مكٻۃ مكٻة كېه مكېە مكېۃ مكېة مکيه مکيە مکيۃ مکية مکیه مکیە مکیۃ مکیة مکېه مکېە مکېۃ مکېة مکٻه مکٻە مکٻۃ مکٻة

2. Pattern Two – apparent stylistic variants

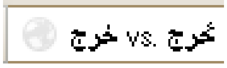
Identical base letters with distinct but closely similar diacritics that are vulnerable to the combination of simplified typography and small sizes. The Latin analogy is, e.g., the ş vs. ș similarity (015F – 0219) Example – GRG: خرج (خ =062E) vs. ځرج (ځ =0681)

Command line on MacOSX



خرج vs. خرج

Command line on Windows



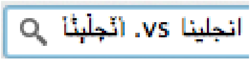
خرج vs. خرج

3. Pattern Three – cool letters

This is the häagen dažš pattern. It can be seen in Arabic messaging, using "cool" Unicode letters for normal text. For certain audiences the diacritics are meaningless, so that they do not interfere with legibility at all, while the underlying Unicode points are totally different.

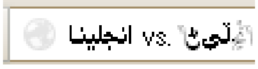
Example 1 – *Angelina* A BGLBBA انجلينا vs. اۡنۡخۡلۡیۡٹۡاۡ

Command line on MacOSX



انجلينا vs. انجلبنا

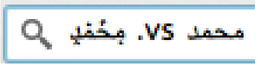
Command line on Windows



انجلينا vs. انجلبٹا

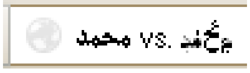
Example 2 – *Mohammed* MGMD محمد vs. مٔحمد

Command line on MacOSX



محمد vs. مخفو

Command line on Windows



محمد vs. مخمد

Additional complications

Keyboards

The *multiple-choice* Unicode characters for certain Arabic graphemes also lead to a vulnerability on the data entry side. For each Arabic-scripted language a virtual keyboard provides access to the relevant repertoire of the Unicode Characters. However, such keyboards use varying Unicode selections for the graphemically identical letters. Yet each of these keyboards can produce passable Arabic, but with underlying encoding that can be totally different. This leads to a practical situation where variant encodings for conceptually identical words are actually produced very easily and inadvertently, with all kinds of unexpected surprises. For example, users trying to enter a URL from a printed source using a regional keyboard, may be not be able to reproduce the intended sequence of code points without being aware of that fact.

Fonts

With the *Pattern One* or K-Y-H group the industry created new, previously non-existent characters. This leads to instability of the underlying encoding that the general target audience cannot perceive. Even the technically savvy audience, who are aware of Unicode, can often not understand or even perceive the differences that the industry tried to encode with this group. As a result Iranian made fonts routinely design final Yeh, whether 064A or 06CC, without dots, regardless the encoding: Persian is not aware of Yeh with dots in final position, therefore in Arabic Yeh is printed without them, too. Arab made fonts for Qurʾānic usage, which, like Persian, never has dots in final position, make the glyphs of 064A identical to those of 06CC.

conclusion

If the goal is to make use of Arabic script in Internationalised Domain Names (IDN) in a secure and robust manner, a solution must be found that addresses the patterns of confusability as described in this document. The patterns of confusability presented here are specific for the Arabic script and its encoding history. On the one hand, more than one encoded character was assigned for functionally equivalent Arabic letters, while on the other hand many new, little known letters were introduced that are often barely distinguishable from the well-known ones. This is exacerbated by the fact that regional keyboards make available arbitrary selections of these characters, so that it is not possible to enter Arabic text in a generic and interchangeable form. Furthermore the encoded characters may be displayed by fonts that do not necessarily adhere to the distinctions made in the encoding. A technical solution for Arabic IDN's must overcome the limitations of fonts, keyboards and duplicate encodings.