Antikythera Publications

Relational Database Design
http://www.AntikytheraPubs.com
jnhummel @ AntikytheraPubs.com

DATABASE DESIGN NOTE SERIES

Exploring Arabic Script Behavior

Multi-script Database Series #6

Prepared by: Frank Oberle (فرينك بيرل) and J. N. Hummel (𝄞 ♩♩♪♪.♩♩)

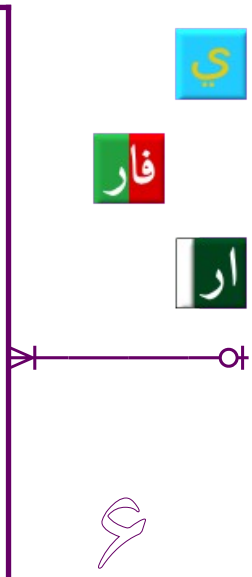
Some examples of Arabic, Devanagari, Hebrew, Korean and Thai Scripts were introduced in Design Note #2 (Exploring Complex Text Layout), but Arabic Script is a bit more complex. While other Scripts consist of characters that may swap positions with their neighbors (e.g. Devanagari), be written right-to-left (e.g. Hebrew), be grouped into blocks (e.g. Korean), or placed above or below their neighbors (e.g. Thai), characters nevertheless remain identifiable as those that were entered.

Arabic Characters, on the other hand, often seem to change as each subsequent character is typed. This can make it difficult for many developers – who may be unfamiliar with such languages – to determine whether the data for which they are responsible is being stored, retrieved, and displayed correctly.

This more detailed look at Arabic Script will help database and application developers make sense of what they are seeing during development and testing of systems required to support any of the more than thirty languages written in Arabic Script, and serve as a basic reference.

Revised for public distribution: 7 September 2018

See page 20 for information on other material from Antikythera Publications.



www.AntikytheraPubs.com

Database Design Note Series on Multi-Language/Multi-Script Databases

1. Exploring Alphabets
2. Exploring Complex Text Layout
3. Exploring UTF-8
4. Evaluating Fonts for use in Multi-Lingual Documents
5. Evaluating Bidirectional Text Entry
6. **Exploring Arabic Script Behavior**

Database Design Note Series – Exploring Arabic Script Behavior

INTRODUCTION

Arabic Script has a strong tradition of decorative and often quite elaborate calligraphy. As a result, its individual characters are often joined (as with English cursive) to prioritize the aesthetics of whole words over individual characters. Based on their position within a word – Initial, Medial (between two others), or Final – characters may take on shapes that differ from their "normal" (Isolate) forms.[1] While current font and rendering technologies are not yet supported widely enough in typical business or personal computing environments to fully reproduce Arabic Script[2], technology practitioners should have a basic understanding of Arabic Script behavior in order to deal with data in the variety of disparate languages that utilize it.

Arabic alphabets are used to write Arwi (in Sri Lanka and Southern India), Azerbaijani (in Iran), Balochi (بلؤچی), Balti, Belarusian (among the Tatars there), Bosniaks (in Bosnia), Brahui (in Pakistan), Central Kurdish (in Iraq and Iran), Chinese (in some areas), Dari (in Pakistan), Fulfulde-Pulaar/Fulani (in Senegal), Hausa, Kashmiri, Kazakh, Kyrgyz (in Central Asia), Luri, some Mandinka dialects (in West Africa), Malay, Mozarabic, Ottoman Turkish, Pashto/Pukhto (پښتو in Afghanistan and Pakistan), Persian/ Farsi (فارسی in Iran), Punjabi (پنجابی in Pakistan – ਪੰਜਾਬੀ in India), Sindhi, Swahili (in East Africa), Urdu (اردو, in Pakistan and India), Uyghur (in China and Central Asia), and Uzbek (in Central Asia) – this is an incomplete list, and some of these are written in multiple Scripts.

For practical reasons, however, this overview will be limited to Modern Standard Arabic, Farsi/Persian, and Urdu. These widely used, yet distinct and unrelated languages[3] should provide a wide enough sampling to demonstrate the basic concepts behind Arabic Script behavior and, more importantly, evaluate the capability of your infrastructure, data handling, and application portfolios to support any of the aforementioned languages.

The keyboard layouts provided in this reference are only a small sampling of those that may be encountered in practice, even for these three languages, but are the ones used for the Arabic typing examples in this series.

The table of glyph forms shows the Isolate form of each character, along with its Initial, Medial and Final forms.

Contents

*

1 It is very important to note however that *only the Isolate forms are stored on disc or transmitted* – usually in UTF-8 form – and it is only during the presentation of these characters on paper or screen that the positional variants are used. Following the preliminaries, a detailed example of how variations of one character are handled is presented on page 18.

2 Due to long standing issues in LibreOffice, for instance, the capability to rotate right-to-left text has simply been removed in current (6.x) releases – and in spite of Unicode, UTF-8, and OpenType, some applications have no support for right-to-left Scripts at all. See Limitations in Current Arabic Script Rendering on page 19 for commentary on even more subtle issues.

3 Modern Standard Arabic is recognized across the Arabic-speaking world in much the same manner as BBC English is understood by English speakers in Boston, New York, Baltimore, Mobile, and Dallas (to say nothing of Toronto, Canberra and Liverpool) whose speech characteristics (and even vocabulary) are quite different. Farsi (Persian) is an Indo-European Language, and closely related to Dari and Tajik, although the latter language is written using Cyrillic Script. Similarly, Urdu (written with Arabic Script) and Hindi (written with Devanagari Script) speakers have few problems talking to each other since these languages have similar origins, differing primarily in their written forms.

تخطيط لوحة المفاتيح العربية

Arabic (Modern Standard) Keyboard Layout

(pronunciation: takhtit lawhat almafatih alearabía) (keystrokes: jo'd' g,pm hglthjdp hguvfdm)

Unicode Arabic Script Planes: 0x0600-0x06ff, 0x0750-0x077f 0x08a0-0x08ff 0xfb50-0xfdff 0xfe70-0xfeff FreeSerif-11 ←RTL

Similar Glyphs / Letter Forms in the Right-to-Left Arabic Abjad Numeric Characters are laid out in left-to-right order.

ى	ي	ئ	ي	ب	ت	ث	ح	ج	خ	س	ش	ص	ض	ط	ظ	ع	غ	ف	ق	ك	ل	ن	ه	ة	لا	لآ	لأ
0649	0620	0626	064a	0628	062a	062b	062d	062c	062e	0633	0634	0635	0636	0637	0638	0639	063a	0641	0642	0643	0644	0646	0647	0629	fefb	fef5	fef7
n		z	d	f	j	e	p	[	o	s	a	w	q	'	/	u	y	t	r	;	g	k	i	m	b	B	G

Note the paired delimiter reversals on the (,) , D , F , C , and V keys.

چیدمان صفحه کلید فارسی

Persian (Farsi) Keyboard Layout

(pronunciation: cheedamon chataykedee farsee) (keystrokes:]dnlhk wtpi ;gdn thvsd)

Unicode Arabic Script Planes: 0x0600-0x06ff, 0x0750-0x077f 0x08a0-0x08ff 0xfb50-0xfdff 0xfe70-0xfeff FreeSerif-12 ←RTL

Similar Glyphs / Letter Forms in the Right-to-Left Persian/Farsi/Arabic Abjad

Numeric Characters are laid out in left-to-right order.

ی	ي	ئ	ي	ب	ت	ث	ح	ج	خ	س	ش	ص	ض	ط	ظ	ع	غ	ف	ق	ك	ل	ن	ه	ة	لا	لآ	لأ
0649	0620	0626	064a	0628	062a	062b	062d	062c	062e	0633	0634	0635	0636	0637	0638	0639	063a	0641	0642	0643	0644	0646	0647	0629	fefb	fef5	fef7
	S			f	j	e	p	[	o	s	a	w	q	x	z	u	y	t	r	Z	g	k	i	J			

Note the paired delimiter reversals on the [(], [)], [{], [}], [[(P)], and [] (O)] keys; Farsi uses French style Guillemets (« and », located on the K and L keys) in place of parentheses. What looks like a stylized comma located on the English “comma key” is actually a an actual alphabetic characterAlso note that the question mark is reversed.

اردو زبان کی بورڈ ترتیب

Urdu (Pakistani, CRULP) Keyboard Layout

(pronunciation: oardoo zabawn see ?? hekseem) (keystrokes: ardw zban ki brwD trtib)

Unicode Arabic Script Planes: 0x0600-0x06ff, 0x0750-0x077f 0x08a0-0x08ff 0xfb50-0xfdff 0xfe70-0xfeff **FreeSerif**-11 ←RTL

Similar Glyphs / Letter Forms in the Right-to-Left Urdu Arabic Abjad — Numeric Characters are laid out in left-to-right order.

ی	ۍ	ئ	ي	ب	ت	ث	ح	ج	خ	س	ش	ص	ض	ط	ظ	ع	غ	ف	ق	ك	ل	ن	ه	ة	لا	لآ	لأ
0649	0620	0626	064a	0628	062a	062b	062d	062c	062e	0633	0634	0635	0636	0637	0638	0639	063a	0641	0642	0643	0644	0646	0647	0629	fefb	fef5	fef7
		U		b	t	C	h	j	K	s	x	S	J	y	V	e	G	f	q		l	n					

The Urdu Language used widely in India and Pakistan is Indic, but written right-to-left using Arabic Script. This phonetic keyboard layout is that defined by the Center for Research in Urdu Language Processing and seems to be the standard layout now used throughout Pakistani government, business, and academic institutions. As Sahar Afshar[4] notes, the Arabic Script used to write Urdu is often spaced more widely than it is when used for other Languages, but this stylistic difference is not implemented by the font itself.

Note the paired delimiter reversals on the [{], [}], [[], and [] keys; Urdu doesn't use "regular/Western" parentheses. Also note that the question mark is reversed.

Actual Arabic forms of the numbers (۱۲۳۴۵۶۷۸۹۰) are given precedence over the "Western" forms but, as with numbers in other right-to-left scripts, are written left-to-right.

4 Several of her lectures and writings are included in References for Further Exploration on page 19.

U.S. English Key Assignments/Codes for Arabic IME Keyboard Layouts

Standard Arabic (qwerty), Persian (Farsi/Iranian), and Urdu (Pakistan, CRULP[5])

Where the English key presses result in the same Unicode Character in any row, those identical characters are shaded in yellow.

KEY	Arabic Standard		Persian/Farsi		Urdu/Pakistani	
A	◌ِ	u+0650	ؤ	u+0624	آ	u+0622
B	لآ	u+fef5	ZWNJ	u+200c[6]	.	u+002e
C	}	u+007d	ژ	u+0698	ث	u+062b
D	]	u+005d	ي	u+064a	ڈ	u+0688
E	◌ُ	u+064f	◌ٍ	u+064d	◌ٰ	u+0670
F	[	u+005b	إ	u+0625	◌ّ	u+0651**
G	لأ	u+fef7	أ	u+0623	غ	u+063a
H	أ	u+0623	آ	u+0622	ھ	u+06be
I	÷	u+00f7	◌ّ	u+0651	◌ِ	u+0650
J	ـ	u+0640	ة	u+0629	ض	u+0636
K	،	u+060c[4]	»	u+00bb	خ	u+062e
L	/	u+002f	«	u+00ab	◌ٔ	u+0654
M	'	u+0027	ء	u+0621	◌ٖ	u+0658
N	آ	u+0622	◌ٔ	u+0654	ں	u+06ba
O	×	u+00d7	]	u+005d	ۃ	u+06c3
P	؛	u+061b	[	u+005b	◌ُ	u+064f
Q	◌َ	u+064e	◌ْ	u+0652	◌ْ	u+0652
R	◌ٌ	u+064c	◌ً	u+064b	ڑ	u+0691
S	◌ٍ	u+064d	ئ	u+0626	ص	u+0635
T	لإ	u+fef9	◌ُ	u+064f	ٹ	u+0679
U	`	u+0060	◌َ	u+064e	ئ	u+0626
V	{	u+007b	◌ٰ	u+0670	ظ	u+0638
W	◌ً	u+064b	◌ٌ	u+064c	◌ّ	u+0651**
X	◌ْ	u+0652	◌ٓ	u+0653	ژ	u+0698
Y	إ	u+0625	◌ِ	u+0650	◌َ	u+064e
Z	~	u+007e	ك	u+0643	ذ	u+0630

Key	Arabic Standard		Persian/Farsi		Urdu/Pakistani	
a	ش	u+0634	ش	u+0634	ا	u+0627
b	لا	u+fefb	ذ	u+0630	ب	u+0628
c	ؤ	u+0624	ز	u+0632	چ	u+0686
d	ي	u+064a	ى	u+06cc	د	u+062f
e	ث	u+062b	ث	u+062b	ع	u+0639
f	ب	u+0628	ب	u+0628	ف	u+0641
g	ل	u+0644	ل	u+0644	گ	u+06af
h	ا	u+0627	ا	u+0627	ح	u+062d
i	ه	u+0647	ه	u+0647	ى	u+06cc
j	ت	u+062a	ت	u+062a	ج	u+062c
k	ن	u+0646	ن	u+0646	ک	u+06a9
l	م	u+0645	م	u+0645	ل	u+0644
m	ة	u+0629	پ	u+067e	م	u+0645
n	ى	u+0649	د	u+062f	ن	u+0646
o	خ	u+062e	خ	u+062e	ہ	u+06c1
p	ح	u+062d	ح	u+062d	پ	u+067e
q	ض	u+0636	ض	u+0636	ق	u+0642
r	ق	u+0642	ق	u+0642	ر	u+0631
s	س	u+0633	س	u+0633	س	u+0633
t	ف	u+0641	ف	u+0641	ت	u+062a
u	ع	u+0639	ع	u+0639	ء	u+0621
v	ر	u+0631	ر	u+0631	ط	u+0637
w	ص	u+0635	ص	u+0635	و	u+0648
x	ء	u+0621	ط	u+0637	ش	u+0634
y	غ	u+063a	غ	u+063a	ے	u+06d2
z	ئ	u+0626	ظ	u+0638	ز	u+0632

5 Center for Research in Urdu Language Processing

6 Zero Width Non Joiner Mark; overrides automatic character/glyph joining behavior.

KEY	Arabic Standard		Persian/Farsi		Urdu/Pakistani	
~	◌ّ	u+0651	÷	u+00f7	◌ً	u+064b
!	!	u+0021	!	u+0021	1	u+0031
@	@	u+0040	’	u+066c[8]	2	u+0032
#	#	u+0023	٬	u+066b	3	u+0033
$	$	u+0024	ریال	u+fdfc[9]	4	u+0034
%	%	u+0025	٪	u+066a	5	u+0035
^	^	u+005e	×	u+00d7	6	u+0036
&	&	u+0026	،	u+060c	7	u+0037
*	*	u+002a	*	u+002a	8	u+0038
(	)	u+0029	)	u+0029	9	u+0039
)	(	u+0028	(	u+0028	0	u+0030
_	_	u+005f	ـ	u+0640	_	u+005f
+	+	u+002b	+	u+002b	+	u+002b
{	<	u+003c	}	u+007d	}	u+007d
}	>	u+003e	{	u+007b	{	u+007b
|	|	u+007c	|	u+007c	|	u+007c
:	:	u+003a	:	u+003a	:	u+003a
"	"	u+0022	؛	u+061b	"	u+0022
<	,	u+002c	>	u+003e	<	u+003c
>	.	u+002e	<	u+003c	٫	u+066b[11]
?	؟	u+061f	؟	u+061f	؟	u+061f
	ریال	u+fdfc +others	ریال	u+fdfc	Rs	u+20a8
B			ZWNJ	u+200c[2]		

Key	Arabic Standard		Persian/Farsi		Urdu/Pakistani	
`	ذ	u+0630	ZWJ	u+200d[7]	~	u+007e
1	1	u+0031	۱	u+06f1	۱	u+06f1
2	2	u+0032	۲	u+06f2	۲	u+06f2
3	3	u+0033	۳	u+06f3	۳	u+06f3
4	4	u+0034	۴	u+06f4	۴	u+06f4
5	5	u+0035	۵	u+06f5	۵	u+06f5
6	6	u+0036	۶	u+06f6	۶	u+06f6
7	7	u+0037	۷	u+06f7	۷	u+06f7
8	8	u+0038	۸	u+06f8	۸	u+06f8
9	9	u+0039	۹	u+06f9	۹	u+06f9
0	0	u+0030	۰	u+06f0	۰	u+06f0
-	-	u+002d	-	u+002d	-	u+002d
=	=	u+003d	=	u+003d	=	u+003d
[	ج	u+062c	ج	u+062c	]	u+005d
]	د	u+062f	چ	u+0686	[	u+005b
\	\	u+005c	\	u+005c	\	u+005c
;	ك	u+0643	ک	u+06a9	؛	u+061b
'	ط	u+0637	گ	u+06af	'	u+0027
,	و	u+0648	و	u+0648	،	u+060c[10]
.	ز	u+0632	.	u+002e	ـ	u+06d4
/	ظ	u+0638	/	u+002f	/	u+002f
		u+20e3[12]				
`			ZWJ	u+200d[3]		

Use of zero-width joiner characters in Persian/Farsi (using Latin script example)

f l y fly or fly? – latter contains u+200c (zwnj) between “f” and “l” to override automatic ligature generation.

7 Zero Width Joiner Mark
8 Arabic Thousands Separator (used with Persian/Farsi)
9 Arabic ligature: Raa’ (ر u+0631) with Farsi Yaa’ (ی u+06cc), Alif (ا u+0627), and Laam (ل u+0644). The word “Riyal”.
10 Arabic Comma (used with Standard Arabic and Urdu)
11 Arabic Decimal Point (used with Urdu)
12 Post-fix Key Cap character used in this table.

Glyph forms for different Character Positions in Arabic Script

Many characters in the Arabic Script, used not only for those Languages illustrated here, but a wide variety of others, may take differing forms depending on their position within a word (i.e. "contextual letter forms"). Arabic script is cursive in nature, and this permits the characters to be connected in a way that mimics – albeit incompletely – traditional Arabic calligraphic practices. The following charts will assist in identifying the characteristics of these forms.

Legend for Arabic Script character form descriptions shown in the charts on the following pages.

Example for the Baa' (ب) character:	Baa'		⇐ Arabic Character Name[13] (transliterated)
Initial Form (بـ) of the Baa' (ب) character ⇒ (glyph truncated when connecting to a following character)	بـ	u+0628 B u+D8A8	⇐ Unicode Hex Value for the Baa' (ب) character in its "normal" Isolate Form, the common English transliteration[14] and its UTF-8 Encoding, also in Hex.
Unicode Hex Value for this Form ⇒	u+fe91		
Medial Form (بـ) of the Baa' (ب) character ⇒ (arm added when connecting between two adjacent characters)	بـ	ب	⇐ If this cell has a full yellow background, it is a "Sun" letter/character.[15] ⇐ Isolate form of this character as it might appear on a keyboard (actual glyph will depend on the keyboard design) and stored on disk and in memory, or transmitted.
Unicode Hex Value for this Form ⇒	u+fe92		
Final Form (ب) of the Baa' (ب) character ⇒ (arm added to Isolate form when used as a final character)	ب	Ar: f Fa: f Ur: b	⇐ Key press to type this character when using the included IME keyboard layouts for standard Arabic, Farsi (Persian), or Urdu (Pakistani) respectively (see pages 4, 5, and 6).
Unicode Hex Value for this Form ⇒	u+fe90		
See detailed example of Form usage on page 18.	B Bird		⇐ Pronunciation w/English example

A user only "types" the normal form of the character, i.e. that which is shown on the top of each physical key cap or virtual keyboard map. Because of the multiple forms a character may take depending on its position within a word, displayed text may change dynamically as each additional character is typed. Not all characters have all forms. Actual "alphabetic characters"[16] are indicated by a double border around the cell containing the Isolate form of the character. The appearance of the characters and glyphs shown in this document may obviously differ depending on the font in use.

In all forms, the number of dot markings above or below the character (in this case "one below") will remain constant, providing a clue to its "real" identity.

13 The names of the characters may differ in Farsi/Persian and Urdu since they are different languages; the Modern Standard Arabic name is used here for convenience.

14 Note that a character with an underscore represents a different sound. See the characters ح (u+062d) and ه (u+0647), for example, pronounced as H and H respectively.

15 Other characters are standard "Moon" characters. The distinction simply – although not *entirely* correctly – means that pronunciation of the definite article "Al" used with nouns is replaced with something more akin to an elision if it is used with a word beginning with a "Moon" character. "The book" (الكتاب) for example is pronounced "al Kitab", while "the sun" (الشّمس) is pronounced "as shum" rather than "al shum." Note that both articles are written as ال ; only the pronunciation differs.

16 The Arabic "alphabets" for most languages are actually Abjads (having only consonants), and "impure Abjads" at that, but such distinctions aren't important here.

	Yaa' w/Hamza above	'Alif w/Hamza below	Waaw w/Hamza above	'Alif w/Hamza above	'Alif w/Madda above	Cutting Hamza	← Character Name
Initial Form	U+0626 U+D8A6	U+0625 U+D8A5	U+0624 U+D8A4	U+0623 U+D8A3	U+0622 U+D8A2	U+0621 ' U+D8A1	← Unicode HexCode ← Transliteration ← UTF-8 Encoding
Medial Form	ئ	إ	ؤ	أ	آ	ء	**BEGIN** ⇦ This Chart, like Arabic text, is displayed right-to-left. See final row for Arabic-specific punctuation characters.
Final Form	Ar: z Fa: S Ur: U	Ar: Y Fa: F ~~Ur:~~	Ar: c Fa: A ~~Ur:~~	Ar: H Fa: G ~~Ur:~~	Ar: N Fa: H Ur: A	Ar: x Fa: M Ur: u	← Arabic Keyboard ← Farsi Keyboard ← Urdu Keyboard ←
	Yaa' is u+06cc	'Alif is u+0627	Waaw is u+0648	'Alif is u+0627	Double 'Alif (u+0627)	Glottal Stop [17]	← Pronunciation

	Haa' (cf u+0647)	Jeem (Geem) (Zheem)	Thaa' (Theh')	Taa' (Teh') Maftuhah	Taa' (Teh') Marbuta	Baa' or Beh'	'Alif
Initial Form	u+fea3 U+062D H U+D8AD	u+fe9f U+062C J U+D8AC	u+fe9b U+062B TH U+D8AB	u+fe97 U+062A T U+D8AA	U+0629 U+D8A9	u+fe91 U+0628 B U+D8A8	U+0627 A U+D8A7
Medial Form	u+fea4 ح	u+fea0 ج	u+fe9c ث	u+fe98 ت	ة	u+fe92 ب	ا
Final Form	u+fea2 Ar: p Fa: p Ur: h	u+fe9e Ar: [Fa: [Ur: j	u+fe9a Ar: e Fa: e Ur: C	u+fe96 Ar: j Fa: j Ur: t	u+fe94 Ar: m Fa: J ~~Ur:~~	u+fe90 Ar: f Fa: f Ur: b	u+fe8e Ar: h Fa: h Ur: a
	H Hit [18]	**ZH** meaSure	**TH**ing [19]	**T** Tap [20]		**B** Bird	**A** Apple (cf u+0649)

17 The cutting Hamza character ء (u+0621) acts as a consonant, but isn't considered an actual alphabetic character. The Cockney pronunciation of t' in bot'tle is similar.
18 The characters ح (u+062d) and ه (u+0647), both forms of the "h" sound, are pronounced as in the emphatic h in **h**it and the softer h in **h**e respectively.
19 The characters ث (u+062b), ذ (u+0630), and ظ (u+0638), all forms of the "th" sound, are pronounced as in **th**ing, **th**at, and **th**ose respectively.
20 The characters ت (u+062a) and ط (u+0637), both forms of the "t" sound, are pronounced as in **t**ap and **t**aught (more emphatic) respectively.

	Sheen (Shin)		Seen (Sin)		Zaa' (Zay')		Raa' (Reh')		Dhaal (Zaal)		Daal		Khaa'	
Initial Form	شـ u+feb7	U+0634 SH U+D8B4	سـ u+feb3	U+0633 S U+D8B3		U+0632 Z U+D8B2		U+0631 R U+D8B1		U+0630 DH U+D8B0		U+062F D U+D8AF	خـ u+fea7	U+062E KH U+D8AE
Medial Form	ـشـ u+feb8	ش	ـسـ u+feb4	س		ز		ر		ذ		د	ـخـ u+fea8	خ
Final Form	ش u+feb6	Ar: a Fa: a Ur: x	س u+feb2	Ar: s Fa: s Ur: s	ز u+feb0	Ar: . Fa: c Ur: z	ر u+feae	Ar: v Fa: v Ur: r	ذ u+feac	Ar: ` Fa: b Ur: Z	د u+feaa	Ar: l Fa: n Ur: d	خ u+fea6	Ar: o Fa: o Ur: K
	SHeep		**S**ip		**Z**oo		rolled R		**TH**at [19]		**D**id [21]		J.S. Ba**CH** (b.1685)	

			Ghayn (Ghain)		Ayn (Ayin)		Dhaa' (Zah')		Taa' (Tah')		Daad (Dad)		Saad (Sad)	
Initial Form		U+063B TO U+063F	غـ u+fecf	U+063A GH U+D8BA	عـ u+fecb	U+0639 c U+D9B9	ظـ u+fec7	U+0638 DH U+D8B8	طـ u+fec3	U+0637 T U+D8B7	ضـ u+febf	U+0636 D U+D8B6	صـ u+febb	U+0635 S U+D8B5
Medial Form		Glyphs from u+63b to u+63f not shown as they are only used for ancient Persian and Azerbaijani	ـغـ u+fed0	غ	ـعـ u+fecc	ع	ـظـ u+fec8	ظ	ـطـ u+fec4	ط	ـضـ u+fec0	ض	ـصـ u+febc	ص
Final Form			غ u+fece	Ar: y Fa: y Ur: G	ع u+feca	Ar: u Fa: u Ur: e	ظ u+fec6	Ar: / Fa: z Ur: V	ط u+fec2	Ar: ' Fa: x Ur: v	ض u+febe	Ar: q Fa: q Ur: J	ص u+feba	Ar: w Fa: w Ur: S
			French **R**, as in **R**ue		Glottal Stop, ≈ ugk		**Th**ose [19]		**T**aught; cf t earlier		**D**aughter [21]		**Sw**ord (open S)	

21 The characters د (Daal u+062f) and ض (Daad u+0636), both forms of the "d" sound, are pronounced like the "d" in **D**id and **D**aughter (a bit more forceful) respectively.

	Noon (Nun)		Meem (Mim)		Laam		Kaaf		Qaaf		Faa' (Feh')		Tatweel (Kashida)	
Initial Form	ن u+fee7	U+0646 N U+D986	م u+fee3	U+0645 M U+D985	ل u+fedf	U+0644 L U+D984	ك[22] u+fedb	U+0643 K U+D983	ق u+fed7	U+0642 Q U+D982	ف u+fed3	U+0641 F U+D981		U+0640 U+D980
Medial Form	نـ u+fee8	نـ	مـ u+fee4	مـ	لـ u+fee0	لـ[23]	كـ u+fedc	كـ	قـ u+fed8	قـ	فـ u+fed4	فـ		ـ
Final Form	ن u+fee6	Ar: k Fa: k Ur: n	م u+fee2	Ar: l Fa: l Ur: m	ل u+fede	Ar: g Fa: g Ur: l	ك u+feda	Ar: ; Fa: Z ~~Ur:~~	ق u+fed6	Ar: r Fa: r Ur: [q]	ف u+fed2	Ar: t Fa: t Ur: f	Under-score ⇒	Ar: J Fa: ❑ ~~Ur:~~
	N Need		**M** Many		**L** Lift (light L)		**K** King		**C** Caught (≈ gk)		**F** Fish		(glyph extension)	

	Kasratan		Dammatan		Fathatan (double FatHa)		Yaa' (Yeh')		'Alif Maksura		Waaw		Haa' (Heh')	
Initial Form		U+064D U+D98D		U+064C U+D98C		U+064B U+D98B	يـ u+fef3	U+064A Y | EE U+D98A	ـد u+fbe8	U+0649 A U+D989		U+0648 W | OO U+D988	هـ u+feeb	U+0647 H U+D987
Medial Form		◌ٍ		◌ٌ		◌ً	يـ u+fef4	يـ	ـد u+fbe9	ى[24]		و[25]	هـ u+feec	هـ
Final Form		Ar: S Fa: Ur:		Ar: R Fa: W ~~Ur:~~		Ar: W Fa: R Ur: ~	ي u+fef2	Ar: d Fa: D ~~Ur:~~	ى u+fef0	Ar: n ~~Fa:~~ ~~Ur:~~	و u+feee	Ar: , Fa: , Ur: w	ه u+feea	Ar: i Fa: i Ur:
	indicates genitive case				Single FatHa is u+064e		**Y**es / Ch**EE**se		Also see u+06cc		**W**onder / F**OO**d		**h** **h**e; cf h note earlier	

22 Note the significant difference between the Isolate form of the Kaaf character and its initial and medial forms.

23 The Laam ل takes on a special shape when followed by an Alif ا alone, with a Madda, or with upper or lower Hamzas. (See u+fef5, u+fef7, u+fef9, and u+fefb).

24 The 'Alif Maksura (ى u+0649) character, used as a word ending 'Alif in certain languages, has all three contextual variants, but there are versions used in other languages which have none, e.g. the Farsi (also Urdu) Yaa' at code point u+06cc. Unicode code point u+feef (ى) is a similar glyph not shown in this document.

25 A FatHa (ً u+064e) is placed over the Waaw (و u+0648) character (وَ) when the latter is used as a consonant.

	Hamza above	’Alif Maddah (above)	Sukoon	Shadda	Vowel: Kasra	Vowel: Damma	Vowel: FatHa
Initial Form	U+0654 U+D994	U+0653 U+D993	U+0652 U+D992	U+0651 U+D991	U+0650 I [26] U+D990	U+064F U [26] U+D98F	U+064E A [26] U+D98E
Medial Form	◌ ء	◌ ـٓ	◌ ه	◌ ـّ	◌ ـِ	◌ ـُ	◌ ـَ
Final Form	~~Ar:~~ Fa: N Ur: L	u+fecf ~~Ar:~~ Fa: X ~~Ur:~~	Ar: X Fa: Q Ur: Q	Ar: ~ Fa: I Ur: W Ur: F	Ar: A Fa: Y Ur: I	Ar: E Fa: T Ur: P	Ar: Q Fa: U Ur: Y
		Doubles an ’Alif	Cancels implied vowel	Doubles a Consonant	Long I, e.g. **B**it	Long U, e.g. **PU**t	Long A, e.g. **PA**t

	Arabic digit One	Arabic digit Zero	Zwarakay	Noon Ghunna	Inverted Damma	Subscript ’Alif	Hamza below
Initial Form	U+0661 1 U+D9A1	U+0660 0 U+D9A0	U+0659 U+D999	U+0658 U+D998	U+0657 U+D997	U+0656 U+D996	U+0655 U+D995
Medial Form	١ [27]	٠ [27]	◌ ـٖ	◌ ـں	◌ ـٗ	◌ ـٰ	◌ ـٕ
Final Form	~~Ar:~~ ~~Fa:~~ ~~Ur:~~	~~Ar:~~ ~~Fa:~~ ~~Ur:~~	~~Ar:~~ ~~Fa:~~ ~~Ur:~~	~~Ar:~~ ~~Fa:~~ Ur: M	~~Ar:~~ ~~Fa:~~ ~~Ur:~~	~~Ar:~~ ~~Fa:~~ ~~Ur:~~	~~Ar:~~ ~~Fa:~~ ~~Ur:~~

26 Corresponding short vowel sounds are seldom written or printed; when the FatHa (ـَ), Damma (ـُ), or Kasra (ـِ) are present, they invariably represent “long” vowel sounds.

27 Two sets of Arabic-Indic digits shown in this list. The first runs from u+0660-0669, and the second from u+06f0-06f9. Most are similar, but the 6, 5, and 4 differ enough (٦ ٤ ٥ versus ۶ ۵ ۴) that a separate sequence, known as Eastern-Arabic-Indic, is used for Farsi and Urdu key mappings shown here. Some sources show the Urdu 4 and 6 differently, though. Note that, unlike Arabic text, all forms of Arabic digits are written from left-to-right when forming number sequences.

	Arabic digit Eight	Arabic digit Seven	Arabic digit Six	Arabic digit Five	Arabic digit Four	Arabic digit Three	Arabic digit Two
Initial Form	U+0668 U+D9A8	U+0667 U+D9A7	U+0666 U+D9A6	U+0665 U+D9A5	U+0664 U+D9A4	U+0663 U+D9A3	U+0662 U+D9A2
Medial Form	٨ [27]	٧ [27]	٦ [27]	٥ [27]	٤ [27]	٣ [27]	٢ [27]
Final Form	~~Ar:~~ ~~Fa:~~ ~~Ur:~~	~~Ar:~~ ~~Fa:~~ ~~Ur:~~	~~Ar:~~ ~~Fa:~~ ~~Ur:~~	~~Ar:~~ ~~Fa:~~ ~~Ur:~~	~~Ar:~~ ~~Fa:~~ ~~Ur:~~	~~Ar:~~ ~~Fa:~~ ~~Ur:~~	~~Ar:~~ ~~Fa:~~ ~~Ur:~~

	Paa’	Tteh	Superscript ’Alif	Arabic Thousands Mark	Arabic Decimal Mark	Arabic Percent Sign	Arabic Digit 9
Initial Form	U+067E U+D9BE	U+0679 U+D9B9	U+0670 U+D9B0	U+066C U+D9AC	U+066B U+D9AB	U+066A U+D9AA	U+0669 U+D9A9
Medial Form	پ	ٹ	◌ٰ	٬ [28]	٫ [28]	٪ [28]	٩ [27]
Final Form	~~Ar:~~ Fa: m Ur: p	~~Ar:~~ ~~Fa:~~ Ur: T	Ar: Fa: V Ur: E	Ar: Fa: Ur:	Ar: Fa: # Ur: >	~~Ar:~~ Fa: % Ur:	~~Ar:~~ ~~Fa:~~ ~~Ur:~~
	P Peel			Numeric Punctuation	Numeric Punctuation	Numeric Punctuation	

28 Some Arabic languages use “lower ASCII” versions (comma, period and % of these numeric punctuation characters depending on context.

		Gaf	Keheh	Jeh	Rreh	Ddal	Tcheem
Initial Form		U+06AF U+DAAF	U+06A9 U+DAA9	U+0698 U+DA98	U+0691 U+DA91	U+0688 U+DA88	U+0686 TCH U+DA86
Medial Form		u+fecf گ u+fecf	ک	ژ	ڑ	ڈ	چ
Final Form		~~Ar:~~ Fa: ' Ur: g u+fecf	~~Ar:~~ Fa: ; Ur: k	~~Ar:~~ Fa: C Ur: X	~~Ar:~~ ~~Fa:~~ Ur: R	~~Ar:~~ ~~Fa:~~ Ur: D	~~Ar:~~ Fa:] Ur: c
							Ch Chip

	Full Stop (period)	Yaa' Baree	Farsi Yaa'	Taa' Marbuta Goal	Heh Goal	Heh Doachashmee	Noon (Nun) Ghunna
Initial Form	U+06D4 U+DB94	U+06D2 U+DB92	U+06CC A U+DB8C	U+06C3 U+DB83	U+06C1 U+DB81	U+06BE U+DABE	U+06BA U+DABA
Medial Form	۔	ے	ی	ۃ [29]	ہ	ھ	ں
Final Form	~~Ar:~~ ~~Fa:~~ Ur: .	~~Ar:~~ ~~Fa:~~ Ur: y u+fecf	~~Ar:~~ Fa: d Ur: i	ة u+fe94 ~~Ar:~~ ~~Fa:~~ Ur: O	~~Ar:~~ ~~Fa:~~ Ur: o	~~Ar:~~ ~~Fa:~~ Ur: H	~~Ar:~~ ~~Fa:~~ Ur: N
	Don't confuse w/u+640		cf Footnote 24 (u+0649)				

29 The Taa' Marbuta ۃ (u+06c3) character used in Urdu is always preceded by a Fatah ٗ (u+064e) character. It is only found at the end of a word.

	Arabic extended digit 6	Arabic extended digit 5	Arabic extended digit 4	Arabic extended digit 3	Arabic extended digit 2	Arabic extended digit 1	Arabic extended digit 0
Initial Form	U+06F6 6 U+DBB6	U+06F5 5 U+DBB5	U+06F4 4 U+DBB4	U+06F3 3 U+DBB3	U+06F2 2 U+DBB2	U+06F1 1 U+DBB1	U+06F0 0 U+DBB0
Medial Form	۶ [27] [30]	۵ [27]	۴ [27] [30]	۳ [27]	۲ [27]	۱ [27]	۰ [27]
Final Form	~~Ar:~~ Fa: 6 **Ur:** 6	~~Ar:~~ Fa: 5 Ur: 5	~~Ar:~~ Fa: 4 **Ur:** 4	~~Ar:~~ Fa: 3 Ur: 3	~~Ar:~~ Fa: 2 Ur: 2	~~Ar:~~ Fa: 1 Ur: 1	~~Ar:~~ Fa: 0 Ur: 0
	cf Digits note above	cf Digits note above	cf Digits note above	cf Digits note above	cf Digits note above	cf Digits note above	cf Digits note above

	Laam + ’Alif	Laam + Hamza below	Laam + Hamza above	Laam + Madda above	Arabic extended digit 9	Arabic extended digit 8	Arabic extended digit 7
Initial Form	U+FEFB EFBBBB	U+FEF9 EFBBB9	U+FEF7 EFBBB7	U+FEF5 EFBBB5[31]	U+06F9 9 U+DBB9	U+06F8 8 U+DBB8	U+06F7 7 U+DBB7
Medial Form	ﻻ	ﻹ	ﻷ	ﷵ	۹ [20]	۸ [20]	۷ [20]
Final Form	Ar: b Fa: Ur:	Ar: I Fa: Ur:	Ar: G Fa: Ur:	Ar: B Fa: Ur:	~~Ar:~~ Fa: 9 Ur: 9	~~Ar:~~ Fa: 8 Ur: 8	~~Ar:~~ Fa: 7 Ur: 7
	Lay g + h	Ligature	Ligature	Ligature	cf Digits note above	cf Digits note above	cf Digits note above

30 Some Urdu keyboards map the “4” key to a character I can’t identify in Unicode; some map the “6” to u+0666. Those shown above seem to be the most common.
31 Character values in the lesser-used higher Arabic Unicode Planes (0xfb50-0xfdff and 0xfe70-0xfeff) require three bytes in UTF-8 format.

	Sheen + 'Alif Maksura	Zero Width Joiner	Zero Width Non-Joiner	Arabic Question Mark	Arabic Semicolon	Arabic Comma
Initial Form	U+FDFC EFB7BC	u+200d zwj E2808D	U+200C zwnj E2808C	U+061F ? U+D89F	U+061B ; U+D89B	U+060C , U+D88C
Medial Form	ريال			؟	؛	،
Final Form	~~Ar:~~ Fa: $ ~~Ur:~~	~~Ar:~~ Fa: ` ~~Ur:~~	~~Ar:~~ Fa: B ~~Ur:~~	Ar: ? Fa: ? Ur: ?	Ar: P Fa: " Ur: ;	Ar: K Fa: & Ur: ,
	Ligature for Riyal [32]	Forces character joins	Prevents character joins	Punctuation	Punctuation	Punctuation

	Post-Fix Key Cap											
Initial Form	U+20E3 ⃣ E283A3			U+20A8 R$ E282A8								
Medial Form	⃣			Rs								
Final Form	u+20e3 is a postfix KeyCap used in this document.			~~Ar:~~ ~~Fa:~~ Ur:								
				Pakistani Rupee								

32 This slightly stylized ligature used to represent the currencies used in Iran, Oman, Qatar, Saudi Arabia, and Yemen, (all called Riyal and spelled ريل , u+631, u+6cc, u+644) only appears on the Farsi/Persian keyboard shown in this document. Many of these countries often use western abbreviations (e.g. SR for the Saudi Riyal) as well.

Arabic Word	← 7	← 6	← 5	← 4	← 3	← 2	← 1
تونس	س	◌ِ	ن	◌ْ	و	◌ُ	ت
	u+0633	u+0650	u+0646	u+0652	u+0648	u+064f	u+062a
	s	A	k	X	,	E	j
Translation	Final Form	Diacritic	Initial Form	Diacritic	Final Form	Diacritic	*Initial Form*
Tunisia *pronounced "tuu-nis"*	س	◌ِ	نـ	◌ْ	ـو	◌ُ	تـ
	u+feb2		u+fee7		u+feee		u+fe97

This page illustrates how different forms of the ت character (see page 8) are displayed using Arabic words for Tunisia, Meter, and Yacht. These three sample words are those used on the web page http://www.myeasyarabic.com/site/arabic_alphabet_taa.htm, since it permits you to hear the actual pronunciation. You can follow this behavior by using an Arabic keyboard layout and pressing [j] [E] [,] [X] [k] [A] and [s] as shown in the table to the left.

The ت isn't affected by typing the diacritic ُ, but as soon as the next alphabetic character و is typed, it changes to its initial form تـ and the و changes to its final form ـو, where the "tail" on the right side of the latter is joined to the truncated left edge of the former.

مترْ				ر	◌ْ	ت	م
				u+0631	u+0652	u+062a	u+0645
				v	X	j	l
Translation				Final Form	Diacritic	*Medial Form*	Initial Form
Meter *pronounced "mitr"*				ـر	◌ْ	ـتـ	مـ
				u+feae		u+fe98	u+fee3

The Arabic word for "meter" illustrates how the same ت character takes on a slightly different (medial) form when it appears between two other alphabetical characters. Type the [l] [j] [X] and [v] characters on an Arabic keyboard. The م is displayed when the [l] is typed, but immediately changes to its initial form مـ once the [j] key is pressed. At this point, our ت appears in its final form ت (since there is no character following it), and remains in that form when the diacritic ْ is typed using the [X] key.

When the [v] is pressed, however, the ر immediately takes on its own final form ـر, while the ت changes from its final form ت to its medial form ـتـ, reflecting that it's now followed by another character.

يَخْت			ت	◌ْ	خ	◌َ	ي
			u+062a	u+0652	u+062e	u+064e	u+064a
			j	X	o	Q	d
Translation			*Final Form*	Diacritic	Medial Form	Diacritic	Initial Form
Yacht *pronounced "yakht"*			ـت	◌ْ	ـخـ	◌َ	يـ
			u+fe96		u+fea8		u+fef3

The word for "yacht" (both Arabic and English forms of this word derive from the early Dutch/low German word *jachtschip*, a fast hunting ship) illustrates the our ت character's final form.

With an Arabic keyboard still selected, type the sequence shown on the left ([d] [Q] [o] [X] and [j]) and carefully observe the displayed sequence for the alterations as each character is typed.

[j]		[X]		[o]		[Q]		[d]	← key typed
يَخْت	←	يَخْ	←	يَخ	←	يَ	←	ي	← display

While by no means a comprehensive test, successful execution of these three examples will confirm the fundamentals are in place for your system to handle Arabic script.

References for Further Exploration

Arabic Script and Related Topics

Arabic Alphabet General Reference [33]
https://en.wikipedia.org/wiki/Arabic_alphabet

Overview of Arabic Calligraphic Styles – Sahar Afshar
https://www.rosettatype.com/blog/2016/05/24/Arabic-calligraphic-styles

World Currency Symbols
https://www.xe.com/symbols.php

Unicode Charts
http://unicode.org/charts/

University of Reading Masters Program: Recommended Reading Material

http://typefacedesign.net/resources/preparation-for-incoming-matd-students/

Limitations in Current Arabic Script Rendering

Unicode Arabic lacks concept of contextually neutral, "inline" characters – Thomas Milo
https://unicode.org/L2/L2014/14109-inline-chars.pdf

Considering the Old; Designing the New (Video 17:29)
"methods of designing Arabic typefaces with a new voice by use of case studies. It will do so by analyzing innovative practice in various aspects of font development such as design, technology, and solutions for harmonization, etc. " – Sahar Afshar
https://www.youtube.com/watch?v=SBCPp-vxMBs

Typo Labs 2018 Lecture Series (includes Making it Fit video below)
https://www.typenetwork.com/news/article/typo-labs-2018-how-far-can-we-go

On Expanding Connections (aka Making it Fit, from TypoLabs 2018) (Video 45:41)
"historic context and use of the Arabic extension, … how, by using variable font technology, the long standing issue of organic extensions can be resolved to make possible a more dynamic typesetting for the Arabic script." – Sahar Afshar and José Miguel Solé Bruning
http://showclipaz.com/at/typo-labs-2018-sahar-afshar-jose-miguel-s_eMYLtwQWjTW4

Related Formal Open Type Proposal from Sahar Afshar and José Miguel Solé Bruning
https://github.com/Microsoft/OpenTypeDesignVariationAxisTags/blob/master/Proposals/Glyph_Extension_Axis/ProposalSummary.md

Possibilities offered by variable font technology for Arabic script (Sahar Afshar Tweet)
https://twitter.com/sahafshar/status/896389627065577472

Language Issues related to Arabic Scripts

The Persian Language (and what makes it fascinating) (Video 8:49)
https://www.youtube.com/watch?v=tZtlDNcbeE8

Comparing Urdu and Hindi Languages 11:52
It may seem that the Hindi (written left-to-right in Devanagari Script) and Urdu (written right-to-left in Arabic Script) Languages are quite different. They are not.
https://www.youtube.com/watch?v=vxSd7p1i_TA

Urdu character names:
http://scriptsource.org/cms/scripts/page.php?item_id=entry_detail&uid=ygnh6uc3nb
Names of Farsi/Persian characters are the same as those used in Modern Standard Arabic and used in this document, although they are pronounced differently.

Typography Issues related to Arabic Scripts (and others)

Liberating Digital Type from the Metal Rectangle (Panel Discussion; Video 57:02)
– Just van Rossum, Bianca Berning, John Hudson, Toshi Omagari, Victor Gaultney, Rob McKaughan, Sahar Afshar
https://www.youtube.com/watch?v=P2JA5EgoDC0
Summary:
https://www.typotalks.com/news/2017/04/08/typo-labs-liberating-digital-type-from-the-metal-rectangle/

Coordinating Styles across differing Scripts – Some Useful Commentary

When combining multiple Scripts in a single document, many of the font matching characteristics used in Latin-only text (terms such as Serif, Sans Serif, Italic, x-height, etc.) have no meaning in other Scripts with which they may be displayed; this is particularly true when mixing Latin and Arabic Scripts. Also see Multi-script Database Design Note #4. Evaluating Fonts for use in Multi-Lingual Documents.

Review of Diodrum Arabic Font – Sahar Afshar on 5 July 2017
Sahar's review concentrates on the stylistic aspects of matching Latin and Arabic Glyphs
https://typographica.org/typeface-reviews/diodrum-arabic/

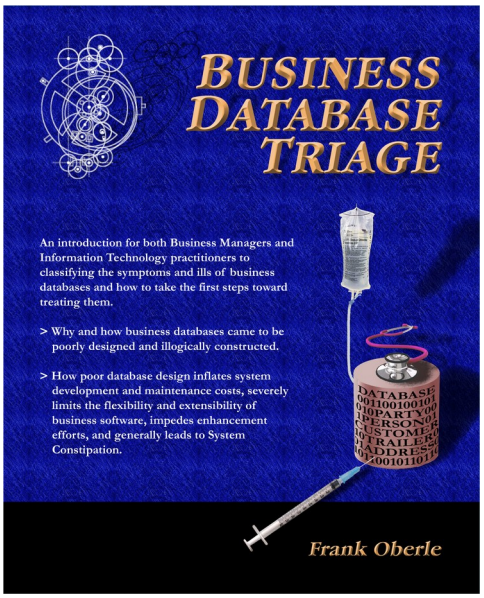
Review of Suisse Int'l Arabic Font – Huda Smitshuijzen AbiFarès on 5 July 2017
See particularly the sizing illustrations with overlays of Latin and Arabic Glyphs
https://typographica.org/typeface-reviews/suisse-intl-arabic/

قف Arabic: r A d Stop

33 A useful reference, although the title should probably be "Arabic Script" rather than "Arabic Alphabet"

Other Publications

Antikythera Publications



More information and sample pages at:
www.AntikytheraPubs.com

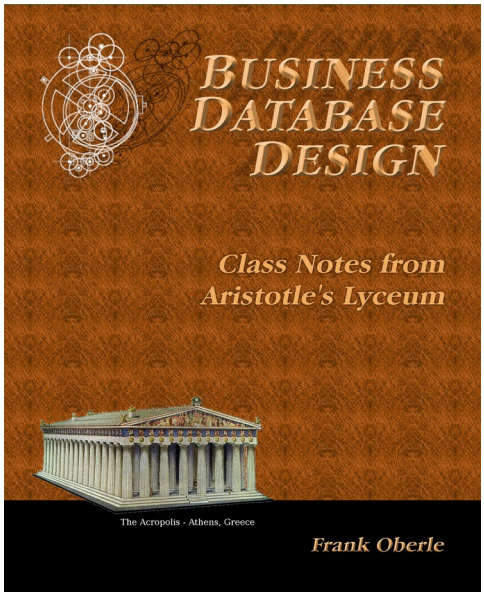
In addition to an ongoing series of Database Design Notes, Antikythera Publications recently released the book "*Business Database Triage*" (ISBN-10: 0615916937) that demonstrates how commonly encountered business database designs often cause significant, although largely unrecognized, difficulties with the development and maintenance of application software. Examples in the book illustrate how some typical database designs impede the ability of software developers to respond to new business opportunities – a key requirement of most businesses.

A number of examples of solutions to curing business system constipation are presented. Urban legends, such as the so-called object-relational impedance mismatch, are debunked – shown to be based mostly on illogical database (and sometimes object) designs.

"*Business Database Triage*" is available through major book retailers in most countries, or from the following on-line vendors, each of which has a full description of the book on their site:

CreateSpace: https://www.createspace.com/4513537

Amazon:
www.amazon.com/Business-Database-Triage-Frank-Oberle/dp/0615916937



A follow-up book, "*Business Database Design – Class Notes from Aristotle's Lyceum*" is due to be available in 2015.

"*Business Database Design*" leads the reader through the logical design and analysis techniques of data organization in more detail than the earlier work – which concentrated more on understanding and identifying problems caused by illogical database design rather than their solutions.

These logical approaches to data organization, espoused by Aristotle and an "A-List" of his successors, have formed the basis for scientific discovery over more than 2,400 years, and directly led to the technology we deal with today, notably including both relational and object theory.

"*Business Database Triage*" explained the reasons why these principles were virtually impossible to apply during the early years of our transition to the use of computers in business, but since the technology is now sufficiently mature that such compromises can no longer be justified, the time has come to relearn logical data organization techniques and apply them to our businesses.

The complete collection of publicly available Database Design Notes, including this series on handling Multi-Language/Multi-Script Databases, is available for free downloading from http://www.antikytherapubs.com/