International Civil Aviation Organization

WORKING PAPER

TAG-MRTD/18-WP/2
22/4/08

Revised
01/05/08

English only

TECHNICAL ADVISORY GROUP ON MACHINE READABLE TRAVEL DOCUMENTS (TAG-MRTD)

EIGHTEENTH MEETING

Montréal, 5 to 8 May 2008

Agenda Item 1: Activities of the NTWG
Agenda Item 1.2 Transliteration Rules (Arabic)

TRANSLITERATION RULES (ARABIC)

Presented by the New Technologies Working Group (NTWG)

1. INTRODUCTION

1.1 The technical report, "Transliteration of Arabic Fonts in Machine Readable Travel Documents", has arisen out of a request by the ICAO Secretariat to WG3/TF3 to undertake research into an Arabic transliteration table for Doc 9303. The need to transliterate Arabic is mainly related to the name fields of the MRTD. Presently names written in the Arabic font are phonetically transcribed to the MRZ in a variety of schemes which lead to much confusion and mismatching of names in databases and other processes.

2. TECHNICAL REPORT SUMMARY

2.1 The MRTD consists of two zones: the Visual Inspection Zone (VIZ) and Machine Readable Zone (MRZ). The representation of the name may be different in each zone, consistent with the usage of these zones. In the VIZ, Doc 9303 mandates that if the name is written in a non-Latin alphabet, then a transliteration into Latin characters must be provided (Part 1, Vol.1, Sect.IV, para.8.3). In the MRZ, the name field is limited to the OCR-B characters 'A' to 'Z' and '<' (Part 1, Vol.1, Sect IV, para.9.2.2), therefore non-Latin characters are not permitted.

2.2 The technical report was compiled and revised to allow interested parties to study the rationale behind the approach and to contribute their ideas and practices.

2.3 The technical report recommends that a phonetic "transcription" be permitted in the VIZ. This may be to a recognised transcription standard or that used by the issuing country or organisation. The TAG/MRTD may wish to standardise the transcription; however, in North Africa mainly French-based transcription schemes are used whereas in the Middle East mainly English-based transcription schemes are used. This tradition may prevent standardisation.

2.4 The technical report recommends that a new "transliteration" scheme be mandated for the MRZ. The aim of this transliteration scheme is to uniquely and reliably convert names written in the Arabic font into the OCR-B Latin characters 'A' to 'Z'. Note that the restriction to the characters 'A' to 'Z' is not a limitation of existing transliteration standards, which means that these schemes cannot be used here.

2.5 At the present revision, the standard Arabic consonants have been appropriately mapped into single Latin characters whenever possible. Where there is a conflict the Latin character 'X' is used as an "escape" character: for example, ه "heh" is mapped to "H; whereas "ح "hah" is mapped to 'XH' in order to preserve its identity. To preserve some of the original structure of the name and keep it recognisable, some Arabic consonants are mapped to compound Latin characters using the escape 'X', for example, ث "theh" is mapped to "XTH". In this way the 'H' in "XTH" is not confused with the 'H' for the Arabic letter ه "heh".

2.6 At the present revision, the optional short vowels ("haracat") are omitted, as are some other signs, for example, ْ "sukun" (absence of a vowel) and ّ "shadda" (doubling of consonant).

2.7 Transliterations are suggested for some other major languages that use the Arabic script, for example, Persian (Farsi), Pashto, Urdu, and the variations found in Moroccan, Tunisian and Maghrib Arabic. Not all these character mappings are complete, and advice is still being sought.

3. KEY AIMS

3.1 The result obtained for the VIZ is to provide a recognizable and traditional phonetic **transcription** of the name.

3.2 The result obtained for the MRZ is to provide a unique representation of the name in the Arabic font, suitable for reliable database and alert list searching. The suggested **transliteration** scheme is reversible, that is, the original name in the Arabic font may be reconstructed from the Latin characters in the MRZ.

3.3 Every effort has been made to make the MRZ representation as recognisable as possible, although an exact match with the phonetic VIZ is not feasible, nor desirable. This is important as the MRZ data will be compared with other data, notably the Passenger Name Record (PNR) held by airlines. It is suggested that once the transliteration rules are incorporated

into Doc 9303 and becomes better known, airlines will start to use it also for the PNR, thus ensuring an exact match without ambiguity.

4. CONCLUSION

4.1 While the transliteration scheme for the MRZ documented in the technical report is not yet complete, the underlying assumption of mapping Arabic characters to the Latin characters 'A' to 'Z' has not been questioned. No other feasible transliteration scheme exists or has been suggested.

5. ACTION BY TAG/MRTD

5.1 The TAG/MRTD is invited to:

a) note the findings and conclusions of the technical report on the "Transliteration of Arabic Fonts in MRTDs";
b) approve the continuation of the work for eventual inclusion of Arabic transliteration rules in table form in the Supplement to Doc 9303; and
c) consider the inclusion of the finished Arabic transliteration table in Doc 9303 as a Normative Appendix.

– – – – – – – – –

Transliteration of Arabic Fonts

in

Machine Readable Travel Documents

TECHNICAL REPORT

Prepared by Mike Ellis (Australia) WG3 for ICAO
Version 2.5, 25-Apr-2008

TABLE OF CONTENTS

DOCUMENTATION HISTORY

Date	Revision	Author	Action
28-July-2007	1.0	Mike Ellis	Initial Draft
23-Sept-2007	1.1	Mike Ellis	Revision for WG3/TF3 Berlin
27-Oct-2007	2.0	Mike Ellis	Revision following WG3 Berlin
17-Jan-2008	2.1	Mike Ellis	Revision following OSCE Workshop Madrid
5-Feb-2008	2.2	Mike Ellis	Revision re comments from Kingdom of Bahrain
15-Feb-2008	2.3	Mike Ellis	Revision following NTWG Christchurch NZ
21-Mar-2008	2.4	Mike Ellis	Revision incorporating comments from Interpol
25-Apr-2008	2.5	Mike Ellis	Revision incorporating comments from Dr Hoogland

1. Scope

The purpose of this Technical Report is to provide guidance and advice to States and to Suppliers regarding the representation of the Arabic font in Latin characters in the Visual Inspection Zone (VIZ) and in the Machine Readable Zone (MRZ) of the Machine Readable Travel Document (MRTD).

2. Introduction

2.1 THE MACHINE READABLE TRAVEL DOCUMENT

The MRTD is defined by ICAO Doc 9303 (ISO/IEC 7501).

The data page of the MRTD consists of two zones:
i) the Visual Inspection Zone (VIZ), which is readable by humans;
ii) the Machine Readable Zone (MRZ), which is readable by machine.

2.2 THE ARABIC FONT

The Arabic font is used by the Arabic language, the official language of about 24 countries from Morocco to Oman. The Arabic font is also used by other languages, notably Farsi in Iran; Pashto and Dari in Afghanistan; Urdu in Pakistan; and many others, including Kurdish, Assyrian, Hausa and Uighur. In the past it was used for the languages of Central Asia, for example, Tajik and Uzbek.

The Arabic font is cursive, and a letter will often change its shape depending upon whether it is standing alone (isolated); at the start of a word (initial); in the body of a word (medial); or at the end (final). For example, the letter ب (beh) changes its shape to بـ at the beginning of the word بكـر (Bakr) - note that Arabic reads from right to left, so the first letter is at the right hand side. We are not concerned here with these different letter shapes (glyphs), only the basic letter code - represented by the isolated shape.

Arabic and the other languages using the Arabic font are usually written using consonants alone. Thus the name **محمد (**Mohammed) as written consists of just four consonants, which may be approximated in Latin as "Mhmd". The vowels are added at the discretion of the translator to achieve a phonetic equivalent. Arabic can also be "vocalized" if the vowel marks ("harakat") are added to modify the pronunciation. However, the harakat are normally omitted.

The standard Arabic font consists of 32 consonants, 18 vowels and dipthongs and 3 other signs. In addition there are over 100 national characters in the Arabic font when used with non-Arabic languages, although some of these are be obsolete and no longer in use.

3. The Arabic font in the MRTD

3.1 VIZ

The VIZ has a mandatory field for the name (fields 6 and 7 of Zone II). In the case of the Machine Readable Passport (MRP), Doc 9303, Part 1, Volume 1, Section IV, Paragraph 8.3 says *"When the mandatory elements of Zones I, II and III are in a national language that does not use the Latin alphabet, a transliteration shall also be provided."*

Thus if the name is written in the Arabic font, a Latin representation shall be included. While Doc 9303 refers to this representation as a "transliteration", it is commonly a phonetic equivalent and should be more correctly termed a "transcription".

For example:

the name[1] in Arabic font: **ابـو بكـر محمـد بـن زكريـا الـرازي**
and a transcription into Latin characters: **Abū Bakr Mohammed ibn Zakarīa al-Rāzi**

Firstly note that paragraph 8.2.3 allows the use of diacritical marks (eg the **ā** in **al-Rāzi**) at the option of the issuing State.

Secondly, note that this particular transcription into Latin characters is only one of many possibilities. The "Database of Arabic Name Variants" website gives the following sixteen variations for **محمد**:

1. Muhammad	2. Moohammad	3. Moohamad
4. Mohammad	5. Mohamad	6. Muhamad
7. Muhamad	8. Mohamed	9. Mohammed
10. Mohemmed	11. Mohemmed	12. Muhemmed
13. Muhamed	14. Muhammed	15. Moohammed
16. Mouhammed		

(see <http://www.kanji.org/cjk/arabic/araborth.htm>).

In some countries it is common to replace the final "d" with "t", so this leads to a total of 32 variations for **محمد** .

The transcription scheme used depends upon the language and regional accent of the Arabic font source (non-Arabic languages such as Farsi, Pashto and Urdu also use the Arabic font); the language of the Latin font speaker; and the transcription scheme used.

[1] Abū Bakr al-Rāzi was a great Persian scientist and doctor of about 1100 years ago. In Persian (Farsi), his name is usually spelt with a final Persian "yeh" (ى), but to avoid confusion we have used the standard Arabic "yeh" (ي).

3.2 MRZ

The Name Field of the MRZ consists, in the case of the MRP, of 39 character positions, and only the OCR-B subset of A-Z and < may be used. Thus Arabic characters shall not be used in the MRZ, and equivalent OCR-B characters must be used to represent them.

It is worth reproducing here the relevant paragraphs of Doc 9303:

9.1 Purpose of the MRZ

9.1.1 MRPs produced in accordance with Doc 9303 Part 1 incorporate an MRZ to facilitate inspection of travel documents and reduce the time taken up in the travel process by administrative procedures. In addition, the MRZ provides verification of the information in the VIZ and may be used to provide search characters for a database inquiry. Equally, it may be used to capture data for registration of arrival and departure or simply to point to an existing record in a database.

9.1.2 ***The MRZ provides a set of essential data elements in a standardized format that can be used by all receiving States regardless of their national script or customs.***

9.1.3 The data in the MRZ are formatted in such a way as to be readable by machines with standard capability worldwide. It must be stressed that the MRZ is reserved for data intended for international use in conformance with international Standards for MRPs. ***The MRZ is a different representation of the data than is found in the VIZ.*** *The VIZ contains data not specifically intended to be read by machine, and herein data can be included in the national script of the issuing State provided that it is also transliterated into Latin-alphabet characters in conformance with 8.3. On the other hand, the constraints posed by machine reading in the MRZ do not permit such flexibility.*

9.2 Properties of the MRZ

9.2.1 In consideration of national privacy laws, the data in the MRZ must be visually readable as well as machine readable. Data presentation must conform to a common standard such that all machine readers configured in conformance with Doc 9303 can recognize each character and communicate in a standard protocol (e.g. ASCII) that is compatible with the technology infrastructure and the processing requirements defined by the receiving State.

9.2.2 To meet these requirements, OCR-B typeface is specified in Doc 9303 as the medium for storage of data in the MRZ. The MRZ as defined herein is recognized as the machine reading technology essential for global interchange and is therefore mandatory in all types of MRPs.

9.3 Constraints of the MRZ

9.3.1 The characters allowed in the MRZ are a common set (as defined in Appendix 8 to this section) which can be used by all States. ***National characters generally appear only in the computer-processing systems of the States in which they apply and are not available globally. They shall not, therefore, appear in the MRZ.***

However, the conversion of the name in the Arabic font to the Latin characters of the MRZ, constrained by the use of only the OCR-B characters A-Z and <, is problematical. In addition, the uncertainty introduced if a phonetic based transcription is allowed means that database searches can become useless.

For example, from the same example used above:

the name in Arabic font: ابـو بكـر محمـد بـن زكريـا الـرازي

and one **transcription** into Latin characters for the MRZ:

`ABU<BAKR<MOHAMMED<IBN<ZAKARIA<AL<RAZI`

However the MRZ is likely to be one of at least 32 variants based on the name “Mohammed” alone. “Zakaria” may be written “Zakariya”; “ibn” as “bin”; and “al” as “el”. Just these variations lead to 256 alternatives.

To draw the contrast, a **transliteration** of the above name محمد , for example, applying the Buckwalter table (see below) to the four Arabic characters, would be “mHmd”. In this case, each Arabic character maps into a single Latin character. No allowance is made for phonetics.

The complete Buckwalter transliteration of the name above is:
Abw<bAkr<mHmd<bn<zkryAY<AlrAzY

Unfortunately, the Buckwalter table uses lower case (a-z) and special characters (‘,|,>,$,<,},*,_,~) so is not suitable for use in the MRZ (see http://www.qamus.org/transliteration.htm)

4. Recommendation for the VIZ

4.1 TRANSCRIPTION IN THE VIZ

As stated above, Doc 9303 in Part 1, Volume 1, Section IV, Paragraph 8.3 mandates the inclusion of a "transliteration" in the VIZ when a national font other than Latin is used for the name.

There is confusion about the terms "transliteration" and "transcription". A "transliteration" is a strictly one-to-one representation of the non-Latin font. A "transcription" is a more loose representation, often based on phonetics (how the name "sounds" when spoken). Of course, often sounds made in one language do not have equivalents in another, and it depends on the target language, for example, "ch", "sh" and "th" are pronounced differently in English and French and German. Compare English "Omar Khayyam" with the German transcription "Omar Chajjam" for the name of the mathematician and poet **عمر خيام.**

There are many "transcription" schemes:

- Deutches Institut für Normung: DIN 31635 (1982)
- Deutsche Morgenländische Gesellschaft (1936)
- International Standards Organisation: ISO/R 233 (1961), ISO 233 (1984)[3], ISO 233-2 (1993)
- British Standards Institute: BS 4280 (1968)
- United Nations Group of Experts on Geographical Names (UNGEGN): UN (1972) [4]
- Qalam (1985)
- American Library Association – Library of Congress: ALA-LC (1997) [1]
- The Encyclopedia of Islam, new edition: EI (1960) [2]

This report recommends that one of these schemes, or one officially used in the country of issue by convention, be used in the VIZ.

4.2 TRANSCRIPTION SCHEMES

Some of the transcription schemes are presented below:

Unicode	Arabic letter	Name[1]	DIN 31635	ISO 233	UN GEGN	ALA-LC	EI
0621	ء	hamza	'	'	'	'	'
0622	آ	alef with madda above	'ā	'â	ā	ā	ā
0627	ا	alef	ā	'			
0628	ب	beh	b	b	b	b	b
0629	ة	teh marbuta	h,t	ṫ	h,t	h,t	a,at
062A	ت	teh	t	t	t	t	t̲
062B	ث	theh	t̲	t̲	th	th	th̲
062C	ج	jeem	ǧ	ǧ	j	j	d̲j̲
062D	ح	hah	ḥ	ḥ	ẖ	ḥ	ḥ
062E	خ	khah	ẖ	h̲	kh	kh	kh̲
062F	د	dal	d	d	d	d	d
0630	ذ	thal	d̲	d̲	dh	dh	dh̲
0631	ر	reh	r	r	r	r	r
0632	ز	zain	z	z	z	z	z
0633	س	seen	s	s	s	s	s
0634	ش	sheen	š	š	sh	sh	sh
0635	ص	sad	ṣ	ṣ	ṣ	ṣ	ṣ
0636	ض	dad	ḍ	ḍ	ḏ	ḍ	ḍ
0637	ط	tah	ṭ	ṭ	ṯ	ṭ	ṭ
0638	ظ	zah	ẓ	ẓ	ẓ̧	ẓ	ẓ
0639	ع	ain	'	'	'	'	'
063A	غ	ghain	ġ	ġ	gh	gh	gh̲
0640	ـ	tatwheel	[graphic filler, not transcrbed]				
0641	ف	feh	f	f	f	f	f
0642	ق	qaf	q	q	q	q	ḳ

[1] The name of the character as given in Unicode and ISO/IEC 10646.

0643	ك	kaf	k	k	k	k	k
0644	ل	lam	l	l	l	l	l
0645	م	meem	m	m	m	m	m
0646	ن	noon	n	n	n	n	n
0647	ه	heh	h	h	h	h	h
0648	و	waw	w	w	w	w	w
0649	ى	alef maksura	ā	ỳ	y	y	ā
064A	ي	yeh	y	y	y	y	y
064B	ً	fathatan	an	á'	a	an	
064C	ٌ	dammatan	un	ú	u	un	
064D	ٍ	kasratan	in	í	i	in	
064E	َ	fatha	a	a	a	a	a
064F	ُ	damma	u	u	u	u	u
0650	ِ	kasra	i	i	i	i	i
0651	ّ	shadda	[double]	¯	[double]	[double]	[double]
0652	ْ	sukun		º			
0670	ٰ	superscript alef	ā	ā	ā	ā	ā

Other national characters are:

067E	پ	peh	p			p	p
0686	چ	tcheh	č			ch,zh	č
0698	ژ	jeh	ž			zh	<u>zh</u>
06A2[1]	ب	feh with dot moved below	f	f		q	
06A4	ڤ	veh	v			v	
06A5	ڤ	feh with 3 dots below	v			v	
06A7[1]	ف	qaf with dot above	q	q		f	
06A8[1]	ق	qaf with 3 dots above	v			v	

[1] Obsolete characters

06AD	ڭ	ng	g			g	g
06AF	گ	gaf	g			g	g

5. Transliteration in the MRZ

5.1 TRANSLITERATION OF EUROPEAN LANGUAGES IN THE MRZ

It is worth considering the situation of the national characters of European languages. Doc 9303 provides a table of "TRANSLITERATIONS RECOMMENDED FOR USE BY STATES" as NORMATIVE APPENDIX 9 to Section IV, Table A. *Transliteration of multinational characters.*

Most of the national characters have their diacritical marks omitted for inclusion in the MRZ. There are a group of nine characters that are treated specially, for example, the character "Ñ" can be transliterated into the MRZ as "NXX", thus preserving its uniqueness and importance for database searches.

For example:

the name in a European national font: **Térèsa CAÑON**

and the transliteration into the MRZ: `CANXXON<<TERESA`

While the MRZ representation appears unaesthetic (and may lead to complaints), the purpose, as stated in the extracts above from Doc 9303, is for machine reading, thus enabling the original name to be recovered for database searches and the like. Thus the MRZ results in the name being recognised as **CAÑON** as distinct from **CANON.**

5.2 USE OF UNICODE

Internally, computers use encoding schemes to represent the characters of different languages. A common encoding scheme is UNICODE, which is nearly equivalent to the ISO/IEC standard 10646 (UNICODE character indices are used in the tables below).

Representations of all the characters of the Arabic font can be found in UNICODE. The UNICODE character indices are usually given as a four digit hexadecimal number (hexadecimal is base 16, and uses the numerals 0-9 and letters A-F to represent the 16 possible numbers).

It has been suggested that UNICODE be used to represent the Arabic font in the MRZ. This has the attraction of having an unique hexadecimal number for each Arabic character. Since all the Arabic characters are located in row 06 (which forms the first two digits of the number), the representation of each Arabic character can be shortened to two hexadecimal numbers.

For example:
ابـو بكـر محمـد بـن زكريـا الـرازي
can be encoded in UNICODE as:

ابـو Alef (ا) - Beh (ب) - Waw (و) => 0627 + 0628 + 0648
بكـر Beh (ب) – Kaf (ك) - Reh (ر) => 0628 + 0643 + 0631
محمد Meem (م) – Hah (ح) – Meem (م) – Dal (د) => 0645 + 062D + 0645 + 062F
بـن Beh (ب) – Noon (ن) => 0628 + 0646
زكريـا Zain (ز) – Kaf (ك) – Reh (ر) – Yeh (ي) – Alef (ا) => 0632 + 0643 + 0631 + 064A + 0627
الـرازي Alef (ا) – Lam (ل) – Reh (ر) – Alef (ا) – Zain (ز) - Yeh (ي) =>
0627 + 0644 + 0631 + 0627 + 0632 + 064A

Dropping the "06" prefix, and using the usual '<' as the separator, the MRZ would be:

272848<284331<452D452F<2846<3243314A27<27443127324A

The main advantage of this scheme is that each Arabic character is uniquely represented, assuming that it is realised in advance that this encoding is UNICODE and the numerals are grouped in pairs. However, there are a number of disadvantages:
a) The standard Doc 9303 does not allow the numerals 0-9 in the name fields of the MRZ. Changing the standard to allow numerals would have immense ramifications for existing machine readers and computer systems;
b) The UNICODE representation is twice as long as the original name, as each Arabic character is transliterated into two UNICODE characters. The original 23 Arabic characters are now 46 UNICODE characters, and even without the separators will not fit in the MRZ; and
c) The UNICODE representation is not in any way readable by humans.

Because of these disadvantages, the use of UNICODE to transliterate Arabic characters in the MRZ is NOT RECOMMENDED.

6. Recommendation for the MRZ

6.1 FACTORS AFFECTING TRANSLITERATION IN THE MRZ

As stated above in Doc 9303 in Part 1, Volume 1, Section IV, Paragraph 9.1.1, "The MRZ provides verification of the information in the VIZ and may be used to provide search characters for a database inquiry." Paragraph 9.1.3 states that "The data in the MRZ are formatted in such a way as to be readable by machines with standard capability worldwide", and "The MRZ is a different representation of the data than is found in the VIZ." However, in paragraph 9.2.1 it is stated that "the data in the MRZ must be visually readable as well as machine readable."

Our aim here is to transliterate the Arabic name into equivalent Latin characters in the MRZ such that there is only one possible representation for the name. This is necessary to avoid ambiguity and make database and alert list searching as accurate as possible. At the same time, the MRZ must be as far as possible a recognizable representation of the name as displayed in the VIZ so that it is visually readable for the purposes of advanced passenger processing and similar uses.

6.2 EXISTING TRANSLITERATION SCHEMES

There are several transliteration schemes in use: Standard Arabic Technical Transliteration System (SATTS), Buckwalter and ASMO 449. These are presented below:

Unicode	Arabic letter	Name	SATTS	Buckwalter	ASMO 449
0621	ء	hamza	E	'	A
0622	آ	alef with madda above	(missing)	|	B
0623	أ	alef with hamza above	(missing)	>	C
0624	ؤ	waw with hamza above	(missing)	&	D
0625	إ	alef with hamza below	(missing)	<	E
0626	ئ	yeh with hamza above	(missing)	}	F
0627	ا	alef	A	A	G
0628	ب	beh	B	b	H
0629	ة	teh marbuta	?	p	I
062A	ت	teh	T	t	J
062B	ث	theh	C	v	K
062C	ج	jeem	J	j	L
062D	ح	hah	H	H	M
062E	خ	khah	O	x	N
062F	د	dal	D	d	O

0630	ذ	thal	Z	*	P
0631	ر	reh	R	r	Q
0632	ز	zain	;	z	R
0633	س	seen	S	s	S
0634	ش	sheen	:	$	T
0635	ص	sad	X	S	U
0636	ض	dad	V	D	V
0637	ط	tah	U	T	W
0638	ظ	zah	Y	Z	X
0639	ع	ain	"	E	Y
063A	غ	ghain	G	g	Z
0640	ـ	tatwheel	(missing)	_	0x60
0641	ف	feh	F	f	a
0642	ق	qaf	Q	q	b
0643	ك	kaf	K	k	c
0644	ل	lam	L	l	d
0645	م	meem	M	m	e
0646	ن	noon	N	n	f
0647	ه	heh	?	h	g
0648	و	waw	W	w	h
0649	ى	alef maksura	(missing)	Y	i
064A	ي	yeh	I	y	j
064B	ً	fathatan	(missing)	F	k
064C	ٌ	dammatan	(missing)	N	l
064D	ٍ	kasratan	(missing)	K	m
064E	َ	fatha	(missing)	a	n
064F	ُ	damma	(missing)	u	o
0650	ِ	kasra	(missing)	i	p
0651	ّ	shadda	(missing)	~	q
0652	ْ	sukun	(missing)	o	r
0670	ٰ	superscript alef	(missing)	`	(missing)

As can be seen from inspection of the tables, these schemes use Latin characters outside of the range A-Z, so are basically unsuitable for use in the MRZ.

The ASMO 449 scheme has an arbitrary allocation of Latin characters, whereas Buckwalter approximates some of the phonetic equivalents.

SATTS does not distinguish between heh (ه) and teh marbuta (ة), or between final yeh (ي) and alif maksura (ى), and it cannot transliterate an alif madda (آ).

6.3 OTHER CONSIDERATIONS

The recommended transliteration scheme cannot be put forward without considering the environment in which the MRTD operates. In particular, the name in the MRZ should be as close as possible in appearance and form as the name derived from other sources. The Passenger Name Record (PNR) used by airlines and forwarded to immigration authorities in Advanced Passenger Information (API) schemes is one example. While the transliteration in the MRZ will almost always not be exactly the same as the transcription in the VIZ (and other phonetic derivatives such as the PNR), the scheme recommended here attempts to make the names in the two zones recognisably similar.

For this purpose the character 'X' is used as an "escape" character in the same sense as in the European National Characters Transliteration table, except only one 'X' is used and it is used before the character it modifies rather than after (eg "XTH" versus "NXX"). One or two characters follow each 'X' to represent one Arabic letter. This use of 'X' is possible as 'X' does not exist in the existing transcription and transliteration schemes for Arabic.

[The difference in the usage of 'X' in Arabic and European transliteration is unlikely to cause confusion. For the proper application of reverse transliteration, the original font must be defined, preferably based on the country of issue.]

In some transliteration entries, a second 'X' is used after the initial 'X': for example, alef with madda above آ is "XAA", alef wasla أ is "XXA". This technique is used primarily to avoid introducing other characters which would make the MRZ less readable by humans.

The intention is that human operators viewing the raw MRZ data from existing systems will be instructed to ignore any 'X' characters. The resulting name should resemble that from other sources. The raw MRZ data will also be lacking vowels that would normally be included in the VIZ transcription and in other sources such as the PNR. However if human operators are instructed that the vowels are missing then the MRZ data should be regarded as a fair representation of the transcribed phonetic version.

The transliteration will also not encompass the assimilation (sandhi) of the article before the "sun letters", and hence the spelling may not match the phonetic transcription of the VIZ (for example, "AL-RAZI" may be "AR-RAZI" in the VIZ). The omission of the "shadda" will also mean that letters doubled in the VIZ representation will only occur as single letters in the MRZ.

6.4 RECOMMENDED TRANSLITERATION SCHEME FOR STANDARD ARABIC

Using the Buckwalter transliteration table as a base, and taking into account the common phonetic equivalents listed in the transcription schemes (paragraph 4.2), a recommended transliteration scheme that only uses the Latin characters A-Z can be formulated. As there is a precedent of using 'X' for variations (paragraph 4.1), the character 'X' is used as an "escape" character to denote that the one or two characters that follow the 'X' represent a single Arabic letter.

This Transliteration Scheme is recommended for use in the MRZ.

Unicode	Arabic letter	Name	Doc 9303	Comments
0621	ء	hamza	XE	
0622	آ	alef with madda above	XAA	6.5.1
0623	أ	alef with hamza above	XAE	6.5.2
0624	ؤ	waw with hamza above	XWE	6.5.3
0625	إ	alef with hamza below	XAI	6.5.4
0626	ئ	yeh with hamza above	XYE	6.5.5
0627	ا	alef	A	
0628	ب	beh	B	
0629	ة	teh marbuta	XTA	6.5.6
062A	ت	teh	T	
062B	ث	theh	XTH	
062C	ج	jeem	J	
062D	ح	hah	XH	6.5.7
062E	خ	khah	XKH	
062F	د	dal	D	
0630	ذ	thal	XDH	
0631	ر	reh	R	
0632	ز	zain	Z	
0633	س	seen	S	
0634	ش	sheen	XSH	
0635	ص	sad	XSS	
0636	ض	dad	XDD	
0637	ط	tah	XTT	
0638	ظ	zah	XZZ	

0639	ع	ain	E	
063A	غ	ghain	G	
0640	ـ	tatwheel	(note 1)	6.5.8
0641	ف	feh	F	
0642	ق	qaf	Q	
0643	ك	kaf	K	
0644	ل	lam	L	
0645	م	meem	M	
0646	ن	noon	N	
0647	ه	heh	H	6.5.7
0648	و	waw	W	
0649	ى	alef maksura	XAY	6.5.9
064A	ي	yeh	Y	
064B	ً	fathatan	(note 1)	6.5.10
064C	ٌ	dammatan	(note 1)	6.5.10
064D	ٍ	kasratan	(note 1)	6.5.10
064E	َ	fatha	(note 1)	6.5.10
064F	ُ	damma	(note 1)	6.5.10
0650	ِ	kasra	(note 1)	6.5.10
0651	ّ	shadda	(note 1)	6.5.11
0652	ْ	sukun	(note 1)	6.5.12
0670	ٰ	superscript alef	(note 1)	6.5.13
0671	ٱ	alef wasla	XXA	6.5.14

The following two letters are commonly used for foreign names:

06A4	ڤ	veh	V	
06A5	ڤ	feh with 3 dots below	XFF	

Note 1: Not encoded.

6.5 COMMENTS ON TRANSLITERATION TABLE

6.5.1 Alef with madda above

Alef with madda above (آ) is not represented in the ALA-LC Romanisation Tables [1]. However, both Interpol [5] and Dr Hoogland [6] recommend the transliteration XAA.

6.5.2 Alef with hamza above

Alef with hamza above (أ) is not represented in the ALA-LC Romanisation Tables [1]. However, Interpol [5] recommends the transliteration XAE and Dr Hoogland [6] recommends XEA.

6.5.3 Waw with hamza above

Waw with hamza above (ؤ) is not represented in the ALA-LC Romanisation Tables [1]. However, Interpol [5] recommends the transliteration XWE and Dr Hoogland [6] recommends XEW.

6.5.4 Alef with hamza below

Alef with hamza below (إ) is not represented in the ALA-LC Romanisation Tables [1]. However, Interpol [5] recommends the transliteration XAI and Dr Hoogland [6] recommends XEI.

6.5.5 Yeh with hamza above

Yeh with hamza above (ئ) is not represented in the ALA-LC Romanisation Tables [1]. However, Interpol [5] recommends the transliteration XYE and Dr Hoogland [6] recommends XEY.

6.5.6 Teh marbuta

Teh marbuta (ة) is represented in the ALA-LC Romanisation Tables [1] as H or T or TAN, depending upon the context. Interpol [5] recommends the transliteration T and Dr Hoogland [6] recommends XTA. We have not used the transliteration T as it would not be compatible with teh (ت), which is also T.

6.5.7 Hah and heh

The transliterations for hah (ح) and heh (ه) have been swapped at the advice of Interpol [5]. Hah is now XH and heh is H.

6.5.8 Tatwheel

Tatwheel (ـ) is a graphic character and not transliterated.

6.5.9 Alef maksura

Alef maksura (ى) is now transliterated as XAY at the recommendation of Dr Hoogland [6]. Other characters are transliterated as XY_, thus the former XY is incompatible.

6.5.10 Short vowels fatha, damma, kasra, fathatan, dammatan and kasratan

The optional short vowels (haracat) are not generally used in names and are not transliterated.

6.5.11 Shadda

Shadda (ّ) is optional and not transliterated. It denotes a doubling of the consonant.

6.5.12 Sukun

Sukun (ْ) denotes the absence of a vowel, is optional, and is not transliterated.

6.5.13 Superscript alef

Superscript alef (ٰ) ("vowel-dagger-alef") is not transliterated.

6.5.14 Alef wasla

Alef wasla (ٱ) is now transliterated as XXA at the recommendation of Interpol [5]. Other characters are transliterated XA_, thus the former XA is incompatible. Dr Hoogland [6] also recommends XXA.

6.5 EXAMPLE OF TRANSLITERATION FOR STANDARD ARABIC

The example above,

ابـو بكـر محمـد بـن زكريـا الـرازي

can be encoded in the MRZ as:

ابـو	Alef (ا) - Beh (ب) - Waw (و) => ABW
بكـر	Beh (ب) – Kaf (ك) - Reh (ر) => BKR
محمد	Meem (م) – Hah (ح) – Meem (م) – Dal (د) => MXHMD
بـن	Beh (ب) – Noon (ن) => BN
زكريـا	Zain (ز) – Kaf (ك) – Reh (ر) – Yeh (ي) – Alef (ا) => ZKRYA
الـرازي	Alef (ا) – Lam (ل) – Reh (ر) – Alef (ا) – Zain (ز) - Yeh (ي) => ALRAZY

ie. ABW<BKR<MXHMD<BN<ZKRYA<ALRAZY

The advantages of this transliteration are:
1. The name in the Arabic font is always transliterated to the same Latin representation. This means that database matches are more likely to result;
2. The process is reversible - the name in the Arabic font can be recovered.

To recover the name in the Arabic font:

ABW	A=Alef (ا) - B=Beh (ب) - W=Waw (و) => **ابـو**
BKR	B=Beh (ب) – K=Kaf (ك) - R=Reh (ر) => **بكـر**
MXHMD	M=Meem (م) – XH=Hah (ح) – M=Meem (م) – D=Dal (د) => **محمد**
BN	B=Beh (ب) – N=Noon (ن) => **بـن**
ZKRYA	Z=Zain (ز) – K=Kaf (ك) – R=Reh (ر) – Y=Yeh (ي) – A=Alef (ا) => **زكريـا**
ALRAZY	A=Alef (ا) – L=Lam (ل) – R=Reh (ر) – A=Alef (ا) – Z=Zain (ز) - Y=Yeh (ي) => **الرازي**

The rationale for omitting the harakat and other diacritical marks is that they are optional and mostly not used. Therefore they should be treated the same way as the diacritical marks on European national characters (eg é, è, ç) which are used for pronunciation purposes.

As well, the optional inclusion of the harakat would be detrimental for accurate database matches.

6.6 RECOMMENDED TRANSLITERATION SCHEME FOR PERSIAN

Persian is spoken in Iran (Farsi), Afghanistan (Dari), Tajikistan and Uzbekistan.

Persian adds four letters to the standard Arabic font:

Unicode	Arabic letter	Name	Doc 9303
067E	پ	peh	P
0686	چ	tcheh	XCH
0698	ژ	jeh	XZH
06AF	گ	gaf	XG

As well, two letters differ in shape:

Unicode	Arabic letter	Name	Doc 9303
06A9	ک	keheh	XKK
06CC	ی	farsi yeh (when in final position)	XAY[1]
06CC	ي	farsi yeh (when in intermediate positions)	Y

The letter "farsi yeh" (ی) is functionally identical to the standard "yeh" (ي) but in the isolated and final forms is graphically identical to the standard "alef maksura" (ى), so it is transliterated as "XY". Database matching algorithms should take this into account.

[1] Dr Hoogland [6]

6.7 RECOMMENDED TRANSLITERATION SCHEME FOR PASHTO

Pashto is spoken in Afghanistan and western Pakistan.

Pashto adds eleven letters to the standard Arabic font:

Unicode	Arabic letter	Name	Doc 9303
067C	ټ	teh with ring	(note 1)
0681	ځ	hah with hamza above	(note 1)
0685	څ	hah with 3 dots above	(note 1)
0689	ډ	dal with ring	(note 1)
0693	ړ	reh with ring	(note 1)
0696	ږ	reh with dot below and dot above	(note 1)
069A	ښ	seen with dot below and dot above	(note 1)
06AB	ګ	kaf with ring	(note 1)
06BC	ڼ	noon with ring	(note 1)
06CD	ۍ	yeh with tail	(note 1)
06D0	ې	E	(note 1)

Note 1: The encodings are undecided in this draft.

6.8 RECOMMENDED TRANSLITERATION SCHEME FOR URDU

Urdu is spoken in Pakistan and India.

Urdu adds seventeen letters to the standard Arabic font. Note that six are the same as Persian (above):

Unicode	Arabic letter	Name	Doc 9303
0679	ٹ	tteh	(note 1)
067E	پ	peh	P
0686	چ	tcheh	C
0688	ڈ	ddal	(note 1)
0691	ڑ	rreh	(note 1)
0698	ژ	jeh	XJ
06A9	ک	keheh	XKK
06AF	گ	gaf	XG
06BA	ں	noon ghunna	(note 1)
06BE	ھ	heh doachashmee	(note 1)
06C0	ہ	heh with yeh above	(note 1)
06C1		heh goal	(note 1)
06C2		heh goal with hamza above	(note 1)
06C3		teh marbuta goal	(note 1)
06CC	ی	farsi yeh	XAY[1]
06D2	ے	yeh barree	(note 1)
06D3	ۓ	yeh barree with hamza above	(note 1)

Note 1: The encodings are undecided in this draft.

[1] Dr Hoogland [6]

6.9 RECOMMENDED TRANSLITERATION SCHEME FOR MOROCCAN, TUNISIAN AND MAGHRIB ARABIC

Moroccan, Tunisian and Maghrib Arabic add four letters to the standard Arabic font:

Unicode	Arabic letter	Name	Doc 9303
069C	ﭽ	seen with 3 dots below and 3 dots above (Moroccan)	(note 1)
06A2	ﭒ	feh with dot moved below (Maghrib)	(note 1)
06A7	ﭪ	qaf with dot above (Maghrib)	(note 1)
06A8	ﭫ	qaf with 3 dots above (Tunisian)	(note 1)

Note 1: These characters are obsolete and not transliterated (at the recommendation of Dr Hoogland [6])

6.10 FURTHER EXAMPLES

Arabic: هاري الشماع
VIZ: **Hari Al-Schamma**
MRZ: HARY<ALXSHMAE

Arabic: سـمير بادمكـدوذيل
VIZ: **Samir Badmakduthal**
MRZ: SMYR<BADMKDWXDHYL

Arabic: جمـال عبـد الناصـر
VIZ: **Gamal Abdel Nasser**
MRZ: JMAL<EBD<ALNAXSSR

Arabic: العبـاس عبـد الله بـن محمـد السـفاح
VIZ: **al-'Abbās 'Abdu'llāh ibn Muhammad as-Saffāh**
MRZ: ALEBAS<EBD<ALLXH<BN<MXHMD<ALSFAXH

Arabic: عبـدالله محمـد بـن عمـر بـن الحسـين فخـر الـدين الـرازي
VIZ: **Abdullah Muhammad ibn Umar ibn al-Husayn Fakhr al-Din al-Razi**
MRZ[1]: EBD<ALLXH<MXHMD<BN<EMR<BN<ALXHSYN<FXKHR

Arabic: عبـدالعزيز بـن متعـب
VIZ: **Abdul Aziz bin Mithab**
MRZ: EBD<ALEZYZ<BN<MTEB

[1] Truncated to 39 characters. The last character is not '<', indicating truncation MAY have occurred.

7. Reverse transliteration of the MRZ

7.1 TABLE FOR REVERSE TRANSLITERATING THE MRZ

Using the table hereunder, the Latin characters in the MRZ can be mapped back into the original Arabic font. Note that 'X' is an "escape" character and the following one or two Latin characters must be used to deduce the corresponding Arabic letter.

MRZ	Name of arabic letter	Arabic letter	Unicode
A	alef	ا	0627
B	beh	ب	0628
C	tcheh (Urdu)	چ	0686
D	dal	د	062F
E	ain	ع	0639
F	feh	ف	0641
G	ghain	غ	063A
H	heh	ه	0647
J	jeem	ج	062C
K	kaf	ك	0643
L	lam	ل	0644
M	meem	م	0645
N	noon	ن	0646
P	peh (Persian, Urdu)	پ	067E
Q	qaf	ق	0642
R	reh	ر	0631
S	seen	س	0633
T	teh	ت	062A
V	veh	ڤ	06A4
W	waw	و	0648
Y	yeh	ي	064A
Z	zain	ز	0632
XAA	alef with madda above	آ	0622
XAE	alef with hamza above	أ	0623
XAI	alef with hamza below	إ	0625
XCH	tcheh (Persian, Urdu)	چ	0686

XDD	dad	ض	0636
XDH	thal	ذ	0630
XE	hamza	ء	0621
XFF	feh with 3 dots below	ڤ	06A5
XG	gaf (Persian, Urdu)	گ	06AF
XH	hah	ح	062D
XJ	jeh (Urdu)	ژ	0698
XKH	khah	خ	062E
XKK	keheh (Persian, Urdu)	ک	06A9
XSH	sheen	ش	0634
XSS	sad	ص	0635
XTA	teh marbuta	ة	0629
XTH	theh	ث	062B
XTT	tah	ط	0637
XWE	waw with hamza above	ؤ	0624
XXA	alef wasla	ٱ	0671
XYE	yeh with hamza above	ئ	0626
XYY	alef maksura or Persian yeh (final position)	ى	0649
XZZ	zah	ظ	0638
XZH	jeh (Persian, Urdu)	ژ	0698

8. References

[1] *ALA-LC Romanization Tables: Transliteration Schemes for Non-Roman Scripts.* Randal K. Berry (ed.). Library of Congress, 1997

[2] *The Encyclopedia of Islam*. New Edition. Leiden, 1960.

[3] *ISO 233:1984. Documentation - Transliteration of Arabic characters into Latin characters.* International Organization for Standardization, 1984-12-15.

[4] *United Nations Romanization Systems for Geographical Names. Report on Their Current Status*. Compiled by the UNGEGN Working Group on Romanization Systems. Version 2.1. June 2002.

[5] *IPSG comments to the document: Transliteration of Arabic Fonts in Machine Readable Travel Documents - Technical Report - Version 2.3 dated 15 Feb 2008*. Interpol, Lyon, 17 March 2008.

[6] Private correspondence, Dr Jan Hoogland, Department of Arabic, University of Nijmegen, the Netherlands, 23 March 2008.

- END -