TAG/MRTD/20-WP/17
6/09/11
English Only

International Civil Aviation Organization

WORKING PAPER

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

TWENTIETH MEETING

Montréal, 7 to 9 September 2011

Agenda Item 2: Activities of the NTWG
Agenda Item 2.9: Transliteration Rules (Arabic)

TRANSLITERATION RULES (ARABIC)

(Presented by the NTWG)

1. INTRODUCTION

1.1 The purpose of the working paper is to inform the TAG of the progress made on the recommended transliteration of Arabic national characters for Doc Doc 9303, *Machine Readable Travel Documents*, Part 1 — *Machine Readable Passports*, Volume1, Section IV, Appendix 9. These recommended transliterations apply to the MRZ only.

2. BACKGROUND

2.1 Transliteration of Arabic characters applies mainly to the name of the holder of the MRTD, and is critical for determining the true identity of the holder. At present, Arabic names are transcribed (phonetically) to Latin characters on an ad-hoc basis and this compromises identity management.

2.2 The only true and reliable source of identification is the original form of the name in Arabic characters. Thus, any transliteration scheme must preserve the original form of the name.

2.3 It would be advantageous to countries using the Arabic script if encoding the name in the MRZ led to machine reading producing the original form of the name in Arabic. This would allow these countries to gain the benefit of machine reading without having to deal with an intermediate form of the name as an inaccurate Latin transcription.

2.4 There is no known existing transliteration scheme for Arabic script that is suitable for the MRZ. All existing schemes are not restricted to ‘A’ to ‘Z’, but use lower case and grammatical signs as well. While such a scheme might be suitable for the Visual Inspection Zone (VIZ), Doc 9303 restricts the name field in the MRZ to ‘A’ to ‘Z’ and ‘<’.

(38 pages)

3. DOC 9303 AND TRANSLITERATION

3.1 Doc 9303 states that if the name in the VIZ is in an alphabet other than Latin, then a transliteration to the Latin alphabet must be provided. Paragraph 8.3 reads as follows:

> "8.3 *Languages and characters*. These specifications provide for entered data in the VIZ to appear in Latin-alphabet characters, i.e. A to Z, and Arabic numerals, i.e. 1234567890. 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…"

3.2 In the case of the Arabic script this transliteration is generally phonetic based and is termed a "transcription". The transcription is usually repeated in the MRZ, with the omission of grammatical signs.

3.3 The MRZ can only contain the OCR-B characters 'A' to 'Z' and '<' in the name field, but, importantly, can be a different representation of the name than that in the VIZ, that is, the two forms need not be identical. Paragraph 9.4.1 reads as follows:

> "9.4.1 Names in the MRZ are represented differently from those in the VIZ. National characters must be transliterated using only the allowed OCR character set *[A..Z]*…"

4. PROPOSED TRANSLITERATION SCHEME FOR THE MRZ

4.1 The Technical Report at Appendix A describes a look-up table approach to transliteration that assigns Latin characters to Arabic characters where a reasonable phonetic match is possible. Some Arabic characters are phonetically similar to others and these are distinguished by the use of 'X' as an "escape" character to differentiate them. The remaining Arabic characters are not phonetically similar to any Latin character and are arbitrarily allocated a Latin character. Implied short vowels that are not present in the original rendition of the name in Arabic are not inserted into the Latin transliteration (unlike the transcription).

4.2 The Technical Report extends the table to Arabic characters that are used in other languages that use the Arabic script (eg Pashto and Farsi).

4.3 The result of using this look-up table process is a mapping of the original name in the Arabic script to an exact Latin representation, and this is a true transliteration. The evidence for this claim is that the original name in Arabic script can be recovered from this Latin transliteration.

5. IMPLEMENTATION ISSUES

5.1 For countries that do not use the Arabic script, the Latin transliteration proposed by this scheme offers the best means of identification of the holder as it is an exact representation of the original name. For countries that have legacy databases of transcribed (phonetic) names, the phonetic name(s) can still be generated from the new transliteration for matching purposes. In fact the matching process should be more reliable as now one phonetic name is being compared with the original name, rather than the historical approach of one phonetic name being matched with another phonetic name, albeit both being derived from the same original Arabic name.

5.2 In general, it is expected that countries that do not use the Arabic script will convert the transliterated name that appears in the MRZ straight to Unicode (ISO/IEC 10646) and store it in that form. In this Unicode form the name from the MRZ will match the Arabic name recorded in DG11, if the MRP is an ePassport.

5.3 It must be noted that the Arabic name, as it appears in the MRZ in the proposed transliterated form, is not generally phonetically readable. Doc 9303 specifically states that the VIZ and the MRZ are different representations of the name and the MRZ is primarily for machine reading. In this case the representation in the MRZ is a direct mapping of the Arabic – and can be in fact read, in Arabic, if the matching Arabic characters are substituted. Paragraph 9.1.2 reads as follows:

> "9.1.2 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 transcribed 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."

5.4 No proposal is made here for the representation of the name in the VIZ. Countries may, if they so desire, continue to provide the phonetic transcription in the VIZ. One country, at least, is known to have a dual birth register with names inscribed in Arabic and phonetic Latin script. Doc 9303, as stated above, makes it mandatory to provide a Latin character equivalent, so it is at the discretion of the issuing state as to whether this is a phonetic transcription, or a copy of the MRZ transliteration. Paragraph 9.3.4 reads as follows:

> "9.3.4 In some instances, names in the MRZ may not appear in the same form as in the VIZ. In the VIZ, non-Latin and national characters may be used to represent more accurately the data in the script of the issuing State or organization."

5.5 Border control officials will need to be trained to differentiate between the VIZ and MRZ forms of the name. It is expected that once the MRZ form is machine read, software can be used to compare the recovered Arabic name with the phonetic form in the VIZ (if so desired).

5.6 Advanced Passenger Information (API) is passenger information, including the name, sent to border control authorities by transport companies (notably airlines) in advance of the travel of the passenger. According to the IATA/CAWG "API Statement of Principles" made at the Facilitation (FAL) Division, Twelfth Session, in Cairo, Egypt, 2004-3-22/4-2, "Required API data should be limited to the data contained in the machine-readable zone of travel documents or obtainable from existing government databases, such as those containing visa issuance information." The IATA/CAWG statement shall be followed and the name, as read from the MRZ, shall be used for API. Subsequently at the border control, the API will match the MRZ data.

5.7 The only other known implementation issue is with airline staff comparing the MRZ form of the name with some other form proffered by the traveler when the travel was purchased, and these two forms may be different. There are two solutions to this problem:

a) Require travelers to record their name in the form shown in the MRZ when purchasing travel; or
b) Modify airline booking systems to record both VIZ and MRZ forms of the name for checking purposes.

5.8 It must be recognized that the overriding requirement is to solve the identity management issue, and that any proposed scheme is bound to have transitional implementation difficulties; however, these are not insolvable and some effort needs to be made to reach this worthwhile goal.

6. ACTION BY THE TAG/MRTD

6.1 The NTWG invites the TAG/MRTD to:

(a) note the work being undertaken in the transliteration of the Arabic script;
(b) provisionally approve the described method, contingent upon the comments by countries that use the Arabic script, specifically for the upcoming ICAO MRTD Regional Seminar, to take place in Qatar in November; and
(c) approve the eventual inclusion of the Arabic transliteration table in Doc 9303, Section IV, Appendix 9.

— — — — — — — —

Transliteration of Arabic Script

in

Machine Readable Travel Documents

TECHNICAL REPORT

Prepared by Mike Ellis (Australia) WG3 for ICAO NTWG

TABLE OF CONTENTS

1. SCOPE 4

2. INTRODUCTION 4

2.1 THE MACHINE READABLE TRAVEL DOCUMENT 4
2.2 THE ARABIC SCRIPT 4

3. THE ARABIC SCRIPT IN THE MRTD 5

3.1 VIZ 5
3.2 MRZ 6

4. RECOMMENDATION FOR THE VIZ 8

4.1 TRANSCRIPTION IN THE VIZ 8
4.2 TRANSCRIPTION SCHEMES 9

5. TRANSLITERATION IN THE MRZ 11

5.1 TRANSLITERATION OF EUROPEAN LANGUAGES IN THE MRZ 11
5.2 USE OF UNICODE 11

6. RECOMMENDATION FOR THE MRZ 12

6.1 FACTORS AFFECTING TRANSLITERATION IN THE MRZ 12
6.2 EXISTING TRANSLITERATION SCHEMES 12
6.3 OTHER CONSIDERATIONS 14
6.4 RECOMMENDED TRANSLITERATION SCHEME FOR STANDARD ARABIC 15
6.5 COMMENTS ON TRANSLITERATION TABLE 17
6.5.1 Alef with madda above 17
6.5.2 Alef with hamza above 17
6.5.3 Waw with hamza above 17
6.5.4 Alef with hamza below 18
6.5.5 Yeh with hamza above 18
6.5.6 Teh marbuta 18
6.5.7 Hah and heh 18
6.5.8 Tatwheel 18
6.5.9 Alef maksura 18
6.5.10 Short vowels fatha, damma, kasra, fathatan, dammatan and kasratan 18
6.5.11 Shadda 19
6.5.12 Sukun 19
6.5.13 Superscript alef 19
6.5.14 Alef wasla 19
6.6 RECOMMENDED TRANSLITERATION SCHEME FOR OTHER LANGUAGES 19
6.7 EXAMPLE OF TRANSLITERATION FOR STANDARD ARABIC 21
6.8 RECOMMENDED TRANSLITERATION SCHEME FOR MOROCCAN, TUNISIAN AND MAGHRIB ARABIC 22
6.9 FURTHER EXAMPLES 23

7. REVERSE TRANSLITERATION OF THE MRZ 25

7.1 TABLE FOR REVERSE TRANSLITERATION OF THE MRZ 25

8. COMPUTER PROGRAMS 28

8.1 ARABIC TO MRZ 28
8.2 MRZ TO ARABIC 29

9. REFERENCES..... 31

APPENDIX 1. COMPLETE TRANSLITERATION TABLE FOR THE MRZ..... 32

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
27-May-2008	2.6	Mike Ellis	Revision following TAG-18 and WG3-37
15-Apr-2011	2.7	Mike Ellis	Revision following NTWG Tokyo
1-Aug-2011	2.8	Mike Ellis	Revision for TAG-20

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 script 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 SCRIPT

The Arabic script is used by the Arabic language, the official language of about 24 countries from Morocco to Oman. The Arabic script 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 script 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 بكر (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 script 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 script consists of 32 consonants, 18 vowels and dipthongs and 3 other signs. In addition there are over 100 national characters in the Arabic script when used with non-Arabic languages, although some of these are be obsolete and no longer in use.

3. The Arabic script 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 script, 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 script: **ابو بكر محمد بن زكريا الرازي**
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**) in the VIZ 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[2] 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		

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 script source (non-Arabic languages such as Farsi, Pashto and Urdu also use the Arabic script); the language of the Latin script 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).
[2] See <http://www.kanji.org/cjk/arabic/araborth.htm>.

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.***

The conversion of the name in the Arabic script 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 script:**ابو بكر محمد بن زكريا الرازي**

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 script 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 script. 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 the English transcription "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]

Some countries maintain their citizens' names in birth or citizen registers in both Arabic and Latin script, where the Latin version is an approved transcription of the Arabic version. These countries may wish to continue to enter the approved Latin transcription in the VIZ.

Recommendation

Doc 9303 in paragraph 8.3, as stated above, makes it mandatory to provide a Latin character equivalent in the VIZ, so it is at the discretion of the Issuing State as to whether this is a phonetic transcription, or a copy of the MRZ transliteration (as described below).

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	T
062B	ث	theh	T	t	th	th	Th
062C	ج	jeem	Ğ	ğ	j	j	Dj
062D	ح	hah	ḥ	ḥ	ẖ	ḥ	ḥ
062E	خ	khah	ḫ	h	kh	kh	Kh
062F	د	dal	D	d	d	d	D
0630	ذ	thal	D	d	dh	dh	Dh
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	ẓ	ẓ	ẓ
0639	ع	ain	'	'	'	'	'
063A	غ	ghain	Ġ	ġ	gh	gh	Gh
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	Zh
06A2[1]	ڢ	feh with dot moved below	f	f		q	
06A4	ڤ	veh	v			v	
06A5	ڦ	feh with 3 dots below	v			v	
06A7[1]	ڧ	qaf with dot above	q	q		f	
06A8[1]	ڨ	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 script: **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 script 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). All Arabic characters are located in row 06 which forms the first two digits of the numbers (ie 06XX).

For example:
ابو بكر محمد بن زكريا الرازي
can be encoded in UNICODE as:

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

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 for reliable identification. 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 fundamentally 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 script 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 ٱ 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"as this is essentially a phonetic feature, 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 "shadda" (symbol to denote doubling of letters) results in the denoted character being repeated in the MRZ (doubled). Search algorithms should take into account that the “shadda” may not always be present.

6.4 RECOMMENDED TRANSLITERATION SCHEME FOR STANDARD ARABIC

Recommendation

Countries should use the table below to transliterate Arabic characters to Latin for the MRZ. This table is repeated in Appendix A to include characters from other Arabic-script based languages (from paragraph 6.6).

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.

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	U	6.5.3

0625	إ	alef with hamza below	I	6.5.4
0626	ئ	yeh with hamza above	XI	6.5.5
0627	ا	alef	A	
0628	ب	beh	B	
0629	ة	teh marbuta	XTA/XAH	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	XDZ	
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	(doubling)	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	پ	feh with 3 dots below	XF	

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. U is used here as *waw with hamza above* is commonly transcribed by "U".

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. The transliteration used here is I as that Latin letter is otherwise unused, and *alef with hamza below* often commences names such as إبراهيم **(Ibrahim)** where the *alef with hamza below* is commonly transcribed by "I".

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. The transliteration used here is XI *as yeh with hamza above* is used in names such as فائز **(Faiz)** where the *yeh with hamza above* is commonly transcribed by "I".

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. The transliteration here of *teh marbuta* has two alternatives: XTA is used generally except if *teh marbuta* occurs at the end of the name component, in which case XAH is used. This is because feminine names often use *teh marbuta* to modify a masculine name, eg فاطمة **(Fatimah)**. Search algorithms should take these two possibilities into account.

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 (ّ) denotes a doubling of the consonant below it, so this is transliterated by doubling the appropriate character. Search algorithms should note that *shaddah* is optional and not sometimes a doubling of the character will be present and sometimes not.

Note the special case of الله (Allah).

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.

6.6 RECOMMENDED TRANSLITERATION SCHEME FOR OTHER LANGUAGES

Persian is spoken in Iran (Farsi), Afghanistan (Dari), Tajikistan and Uzbekistan.
Pashto is spoken in Afghanistan and western Pakistan.
Urdu is spoken in Pakistan and India.

Unicode	Arabic letter	Language	Name	Doc 9303
0679	ٹ	Urdu	Tteh	XXT
067E	پ	Persian, Urdu	Peh	P
067C	ټ	Pashto	teh with ring	XRT
0681	ځ	Pashto	hah with hamza above	XKE
0685	څ	Pashto	hah with 3 dots above	XXH
0686	چ	Persian, Urdu	Tcheh	XC
0688	ڈ	Urdu	Ddal	XXD
0689	ډ	Pashto	dal with ring	XDR
0691	ڑ	Urdu	Rreh	XXR

0693	ڕ	Pashto	reh with ring	XRR
0696	ڙ	Pashto	reh with dot below and dot above	XRX
0698	ژ	Persian, Urdu	Jeh	XJ
069A	ښ	Pashto	seen with dot below and dot above	XXS
06A9	ک	Persian, Urdu	keheh	XKK
06AB	گ	Pashto	kaf with ring	XXK
06AD	ڭ		Ng	XNG
06AF	گ	Persian, Urdu	gaf	XGG
06BA	ں	Urdu	noon ghunna	XNN
06BC	ڼ	Pashto	noon with ring	XXN
06BE	ھ	Urdu	heh doachashmee	XDO
06C0	ۀ	Urdu	heh with yeh above	XYH
06C1		Urdu	heh goal	XXG
06C2		Urdu	heh goal with hamza above	XGE
06C3		Urdu	teh marbuta goal	XTG
06CC	ی	Persian, Urdu	farsi yeh	XYA[3]
06CD	ۍ	Pashto	yeh with tail	XXY
06D0	ې	Pashto	Yeh	Y[4]
06D2	ے	Urdu	Yeh barree	XYB
06D3	ۓ	Urdu	yeh barree with hamza above	XBE

[3] 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 could be transliterated as 'Y' or "XAY". Database matching algorithms should take this into account.

[4] The character "Pashto yeh" (ې) is functionally identical to the standard "yeh" (ي).

6.7 EXAMPLE OF TRANSLITERATION FOR STANDARD ARABIC

The example above,
ابو بكر محمد بن زكريا الرازي
can be encoded in the MRZ as:

ابو	Alef (ا) - Beh (**ب**) - Waw (**و**) => ABW
بكر	Beh (**ب**) – Kaf (**ك**) - Reh (**ر**) => BKR
محمد	Meem (**م**) – Hah (**ح**) – Meem (**م**) – Dal (**د**) => MXHMD
بن	Beh (**ب**) – Noon (**ن**) => BN
زكريا	Zain (**ز**) – Kaf (**ك**) – Reh (**ر**) – Yeh (**ي**) – Alef (ا) => ZKRYA
الرازي	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 script 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 script can be recovered.

To recover the name in the Arabic script:

ABW	A=Alef (ا) - B=Beh (**ب**) - W=Waw (**و**) => **ابو**
BKR	B=Beh (**ب**) – K=Kaf (**ك**) - R=Reh (**ر**) => **بكر**
MXHMD	M=Meem (**م**) – XH=Hah (**ح**) – M=Meem (**م**) – D=Dal (**د**) => **محمد**
BN	B=Beh (**ب**) – N=Noon (**ن**) => **بن**
ZKRYA	Z=Zain (**ز**) – K=Kaf (**ك**) – R=Reh (**ر**) – Y=Yeh (**ي**) – A=Alef (ا) => **زكريا**
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.

6.8 RECOMMENDED TRANSLITERATION SCHEME FOR MOROCCAN, TUNISIAN AND MAGHRIB ARABIC

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

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])

6.9 FURTHER EXAMPLES

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

Arabic: **سمير بادمكدوذيل**
VIZ: **Samir Badmakduthal**
MRZ: SMYR<BADMKDWXDHYL<<<<<<<<<<<<<<<<<<<<<<

Arabic: **جمال عبد الناصر**
VIZ: **Gamal Abdel Nasser**
MRZ: JMAL<EBD<ALNAXSSR<<<<<<<<<<<<<<<<<<<<<<

Arabic: **العباس عبد الله بن محمد السفاح**
VIZ: **al-'Abbās 'Abdu'llāh ibn Muhammad as-Saffāh**
MRZ: ALEBAS<EBD<ALLXH<BN<MXHMD<ALSFAXH<<<<<<

Arabic: **عبدالله محمد بن عمر بن الحسين فخر الدين الرازي**
VIZ: **Abdullah Muhammad ibn Umar ibn al-Husayn Fakhr al-Din al-Razi**
MRZ[1]: EBD<ALLXH<MXHMD<BN<EMR<BN<ALXHSYN<FXKHR

Arabic: **عبدالعزيز بن متعب**
VIZ: **Abdul Aziz bin Mithab**
MRZ: EBD<ALEZYZ<BN<MTEB<<<<<<<<<<<<<<<<<<<<

Arabic: إسماعيل عزّ الدين
VIZ: **Isma'il Izz-ud-din**
MRZ: ISMAEYL<EZZ<ALDYN<<<<<<<<<<<<<<<<<<<<

Arabic: جميلة نعيمة
VIZ: **Jamillah Na'ima**
MRZ: JMYLXAH<NEYMXAH<<<<<<<<<<<<<<<<<<<<<<

6.10 ORDER OF NAMES IN THE MRZ

Doc 9303 states:

> *9.6 The name consists of primary and secondary identifiers which shall be separated by two filler characters (<<). Components within the primary or secondary identifiers shall be separated by a single filler character (<).*

As the primary purpose of this Technical Report is transliteration, no attempt has been made to define primary and secondary identifiers in Arabic names. It is for the issuing authority to make that determination. But as an example:

the name in Arabic script:**ابو بكر محمد بن زكريا الرازي**

1) if the component BN<ZKRYA<ALRAZY is considered the primary identifier, then the MRZ is:

```
BN<ZKRYA<ALRAZY<<ABW<BKR<MXHMD<<<<<<<<<
```

2) if the component ALRAZY is considered the primary identifier, then the MRZ is:

```
ALRAZY<<ABW<BKR<MXHMD<BN<ZKRYA<<<<<<<<<
```

7. Reverse Transliteration of the MRZ

7.1 TABLE FOR REVERSE TRANSLITERATION OF THE MRZ

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

MRZ	Name of arabic letter	Arabic letter	Unicode
A	Alef	ا	0627
B	Beh	ب	0628
D	Dal	د	062F
E	Ain	ع	0639
F	Feh	ف	0641
G	Ghain	غ	063A
H	Heh	ه	0647
I	Alef with hamza below	إ	0625
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
U	Waw with hamza above	ؤ	0624
V	Veh	ڤ	06A4
W	Waw	و	0648
Y	Yeh or Yeh (Pashto)	ي / ې	064A/06D0
Z	Zain	ز	0632
XAA	Alef with madda above	آ	0622
XAE	Alef with hamza above	أ	0623
XAH	Teh marbuta (see also XTA)	ة	0629

XAY	Alef maksura	ى	0649
XBE	Yeh barree with hamza above	ۓ	06D3
XC	Tcheh (Persian, Urdu)	چ	0686
XDH	Thal	ذ	0630
XDO	Heh doachashmee	ھ	06BE
XDR	Dal with ring (Pashto)	ڍ	0689
XDZ	Dad	ض	0636
XE	Hamza	ء	0621
XF	Feh with 3 dots below	ڤ	06A5
XGG	Gaf (Persian, Urdu)	گ	06AF
XGE	Heh goal with hamza above (Urdu)		06C2
XH	Hah	ح	062D
XI	Yeh with hamza above	ئ	0626
XJ	Jeh (Urdu)	ژ	0698
XKE	Hah with hamza above (Pashto)	ځ	0681
XKH	Khah	خ	062E
XKK	Keheh (Persian, Urdu)	ک	06A9
XNN	Noon ghunna (Urdu)	ں	06BA
XNG	Ng	ڭ	06AD
XRR	Reh with ring (Pashto)	ڑ	0693
XRT	Teh with ring	ټ	067C
XRX	Reh with dot below and dot above (Pashto)	ږ	0696
XSH	Sheen	ش	0634
XSS	Sad	ص	0635
XTA	Teh marbuta (see also XAH)	ة	0629
XTG	Teh marbuta goal (Urdu)		06C3
XTH	Theh	ث	062B
XTT	Tah	ط	0637
XXA	Alef wasla	ٱ	0671
XXD	Ddal (Urdu)	ڈ	0688
XXG	Heh goal (Urdu)		06C1
XXH	Hah with 3 dots above (Pashto)	څ	0685
XXK	Kaf with ring (Pashto)	ګ	06AB

XXN	Noon with ring (Pashto)	ڼ	06BC
XXR	Rreh (Urdu)	ڑ	0691
XXS	Seen with dot below and dot above (Pashto)	ښ	069A
XXT	Tteh (Urdu)	ټ	0679
XXY	Yeh with tail (Pashto)	ۍ	06CD
XYA	Farsi yeh (Persian, Urdu)	ی	06CC
XYB	Yeh barree (Urdu)	ے	06D2
XYH	Heh with yeh above (Urdu)	ۀ	06C0
XZZ	Zah	ظ	0638

8. Computer Programs

8.1 ARABIC TO MRZ

This program written in Python is offered as an example of converting Arabic characters (in Unicode) to the MRZ format.

The Arabic characters are contained in a file “Arabic source.txt” and the corresponding MRZ data is written to a file “MRZ output.txt”.

```
**************************************************************************

#    # -*- coding: iso-8859-15 -*-

import unicodedata
import encodings.utf_8_sig
import codecs

# TRANSLITERATE
def Arabic_to_MRZ(unicode_string):
   transform = {0x20: '<', 0x21: 'XE', 0x22: 'XAA', 0x23: 'XAE', 0x24: 'U',
               0x25: 'I', 0x26: 'XI', 0x27: 'A', 0x28: 'B', 0x29: 'XAH',
               0x2A: 'T', 0x2B: 'XTH', 0x2C: 'J', 0x2D: 'XH', 0x2E: 'XKH',
               0x2F: 'D', 0x30: 'XDH', 0x31: 'R', 0x32: 'Z', 0x33: 'S', 0x34: 'XSH',
               0x35: 'XSS', 0x36: 'XDZ', 0x37: 'XTT', 0x38: 'XZZ', 0x39: 'E',
               0x3A: 'G', 0x41: 'F', 0x42: 'Q', 0x43: 'K', 0x44: 'L',
               0x45: 'M', 0x46: 'N', 0x47: 'H', 0x48: 'W', 0x49: 'XAY',
               0x4A: 'Y', 0x71: 'XXA', 0x79: 'XXT', 0x7E: 'P', 0x7C: 'XRT',
               0x81: 'XKE', 0x85: 'XXH', 0x86: 'XC', 0x88: 'XXD', 0x89: 'XDR',
               0x91: 'XXR', 0x93: 'XRR', 0x96: 'XRX', 0x98: 'XJ', 0x9A: 'XXS',
               0xA4: 'XV', 0xA5: 'XF', 0xA9: 'XKK', 0xAB: 'XXK', 0xAD: 'XNG',
               0xAF: 'XGG', 0xBA: 'XNN', 0xBC: 'XXN', 0xBE: 'XDO', 0xC0: 'XYH',
               0xC1: 'XXG', 0xC2: 'XGE', 0xC3: 'XTG',
               0xCC: 'XYA', 0xCD: 'XXY', 0xD0: 'Y', 0xD2: 'XYB', 0xD3: 'XBE'}
   name_in = unicode_string
   name_out = ""
   for c in name_in:
# check for shadda (double)
      if ord(c) == 0x51:
         name_out = name_out + char
      else:
         if ord(c) in transform:
            char = transform[ord(c)]
            name_out = name_out + char
   print name_out
   return name_out

#
#   MAIN - Arabic to MRZ
#

# open input and output files

fin = encodings.utf_8_sig.codecs.open('Arabic source.txt', 'r') #b', 'utf-8-sig', 'ignore', 1)
fout = open('MRZ output.txt', 'w')

# loop through the input file
```

```
try:
   for arabic_name in fin:
      MRZ_name = Arabic_to_MRZ(arabic_name)
      fout.write(MRZ_name)
      fout.write('\n')
finally:
   fin.close()
fout.flush()
fout.close()
```

****

8.2 MRZ TO ARABIC

This program written in Python is offered as an example of converting MRZ characters to Arabic characters (in Unicode).

The MRZ characters are contained in a file “MRZ source.txt” and the corresponding Arabic data is written to a file “Arabic output.txt”.

****

```
#    # -*- coding: iso-8859-15 -*-

import unicodedata
import encodings.utf_8_sig
import codecs

# TRANSLITERATE
def MRZ_to_Arabic(ascii_string):
   transform = { '<': 0x20, 'XE': 0x21, 'XAA':0x22, 'XAE': 0x23, 'U': 0x24,
               'I': 0x25, 'XI': 0x26, 'A': 0x27, 'B': 0x28, 'XAH': 0x29,
               'T': 0x2A, 'XTH': 0x2B, 'J': 0x2C, 'XH': 0x2D, 'XKH': 0x2E,
               'D': 0x2F, 'XDH': 0x30, 'R': 0x31, 'Z': 0x32, 'S': 0x33, 'XSH': 0x34,
               'XSS': 0x35, 'XDZ': 0x36, 'XTT': 0x37, 'XZZ': 0x38, 'E': 0x39,
               'G': 0x3A, 'F': 0x41, 'Q': 0x42, 'K': 0x43, 'L': 0x44, 'M': 0x45,
               'N': 0x46, 'H': 0x47, 'W': 0x48, 'XAY': 0x49, 'Y': 0x4A, 'XXA': 0x71,
               'XXT': 0x79, 'P': 0x7E, 'XRT': 0x7C, 'XKE': 0x81, 'XXH': 0x85,
               'XC': 0x86, 'XXD': 0x88, 'XDR': 0x89, 'XXR': 0x91, 'XRR': 0x93,
               'XRX': 0x96, 'XJ': 0x98, 'XXS': 0x9A, 'XV': 0xA4, 'XF': 0xA5,
               'XKK': 0xA9, 'XXK': 0xAB, 'XNG': 0xAD, 'XGG': 0xAF,
               'XNN': 0xBA, 'XXN': 0xBC, 'XDO': 0xBE, 'XYH': 0xC0,
               'XXG': 0xC1, 'XGE': 0xC2, 'XTA': 0x29, 'XTG': 0xC3, 'XYA': 0xCC,
               'XXY': 0xCD, 'I': 0xD0, 'XYB': 0xD2, 'XBE': 0xD3}
   name_in = ascii_string
   name_out = ""
   # if this character is not X, does it appear by itself in the table?
   search_string = ''
   last_string = ''
   iloop = 0
   while iloop < len(name_in):
      search_string = search_string + name_in[iloop]
      if search_string in transform:
         if search_string <> last_string:
            name_out = name_out + chr((transform[search_string]))
         #insert shadda if double found
```

```
            else:
                name_out = name_out + chr(0x51)
            if search_string <> '<':
                name_out = name_out + chr(0x06)
            else:
                name_out = name_out + chr(0x00)
            #remember last string
            if search_string <> '<':
                last_string = search_string
            else:
                last_string = ''
            #clear the search string once found
            search_string = ''
        iloop = iloop + 1
    print name_out
    return name_out

#
#   MAIN - MRZ to Arabic
#

# open input and output files

fin = open('MRZ source.txt', 'r')
fout = open('Arabic output.txt', 'wb') #b', 'utf-8-sig', 'strict', 1)
fout.write(encodings.utf_8_sig.codecs.BOM)


# loop through the input file

try:
    for MRZ_name in fin:
        Arabic_name = MRZ_to_Arabic(MRZ_name)
        Arabic_name = Arabic_name + chr(0x0D) + chr(0x00) + chr(0x0A) + chr(0x00)
        fout.write(Arabic_name)
finally:
    fin.close()
fout.flush()
fout.close()

*********************************************************************************
```

9. 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.

[7] *Comments on the Translation of Arabic Fonts in Machine Readable Travel Documents TECHNICAL REPORT AMA 13052008,* Mr Abdalla M. Askar, Emirates Identity Authority.

Appendix 1. Complete Transliteration Table for the MRZ

Unicode	Arabic letter	Name	MRZ
0621	ء	hamza	XE
0622	آ	alef with madda above	XAA
0623	أ	alef with hamza above	XAE
0624	ؤ	waw with hamza above	U
0625	إ	alef with hamza below	I
0626	ئ	yeh with hamza above	XI
0627	ا	alef	A
0628	ب	beh	B
0629	ة	teh marbuta	XTA/XAH[5]
062A	ت	teh	T
062B	ث	theh	XTH
062C	ج	jeem	J
062D	ح	hah	XH
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	XDZ
0637	ط	tah	XTT
0638	ظ	zah	XZZ
0639	ع	ain	E
063A	غ	ghain	G
0640	ـ	tatwheel	
0641	ف	feh	F

[5] XTA is used generally except if *teh marbuta* occurs at the end of the name component, in which case XAH is used.

0642	ق	qaf	Q
0643	ك	kaf	K
0644	ل	lam	L
0645	م	meem	M
0646	ن	noon	N
0647	ه	heh	H
0648	و	waw	W
0649	ى	alef maksura	XAY
064A	ي	yeh	Y
064B	ً	fathatan	
064C	ٌ	dammatan	
064D	ٍ	kasratan	
064E	َ	fatha	
064F	ُ	damma	
0650	ِ	kasra	
0651	ّ	shadda	[DOUBLE][6]
0652	ْ	sukun	
0670	ٰ	superscript alef	
0671	ٱ	alef wasla	XXA
0679	ٹ	Tteh	XXT
067E	پ	Peh	P
067C	ټ	teh with ring	XRT
0681	ځ	hah with hamza above	XKE
0685	ڿ	hah with 3 dots above	XXH
0686	چ	Tcheh	XC
0688	ڈ	Ddal	XXD
0689	ډ	dal with ring	XDR
0691	ڑ	Rreh	XXR
0693	ړ	reh with ring	XRR
0696	ږ	reh with dot below and dot above	XRX
0698	ژ	Jeh	XJ

[6] Shadda denotes doubling: Latin character or sequence is repeated eg عبّاس becomes E**BB**AS; فضّة becomes F**XDZXDZ**XAH.

069A	ښ	seen with dot below and dot above	XXS
069C	ڜ	seen with 3 dots below and 3 dots above	
06A2	ڢ	feh with dot moved below	
06A7	ڧ	qaf with dot above	
06A8	ڨ	qaf with 3 dots above	
06A9	ک	keheh	XKK
06AB	ګ	kaf with ring	XXK
06AD	ڭ	Ng	XNG
06AF	گ	gaf	XGG
06BA	ں	noon ghunna	XNN
06BC	ڼ	noon with ring	XXN
06BE	ھ	heh doachashmee	XDO
06C0	ۀ	heh with yeh above	XYH
06C1		heh goal	XXG
06C2		heh goal with hamza above	XGE
06C3		teh marbuta goal	XTG
06CC	ی	farsi yeh	XYA
06CD	ۍ	yeh with tail	XXY
06D0	ې	Yeh	Y
06D2	ے	Yeh barree	XYB
06D3	ۓ	yeh barree with hamza above	XBE