International Civil Aviation Organization

INFORMATION PAPER

TAG/MRTD/20-IP/4
22/08/11
English Only

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 2.13: Transliteration of Arabic Characters in MRTDS

TRANSLITERATION OF ARABIC CHARACTERS IN MRTDS

(Presented by NTWG)

1. INTRODUCTION

1.1 The accurate representation of the name in machine readable travel documents (MRTDs) is required for proper identification of the holder of the document. ICAO Doc 9303 requires all non-Latin names to be transliterated into Latin. The transliteration of the name in Arabic characters has taken the form of phonetic transcription, where the nearest sounding representation in Latin characters is used. Not only is this undesirable because of the amount of variation and uncertainty introduced, but also this does not preserve the original Arabic name, a situation that is not respectful of Arabic cultural heritage or of the status of the holder.

1.2 The adoption of the recommended practice described below for Arabic transliteration in Doc 9303 will achieve accurate representation of Arabic names as they will no longer be subject to the variations of phonetic transcription. There will be a transition period where existing phonetic transcriptions will coexist with the new exact transliterations in databases. However, this is not an insurmountable problem, as for matching purposes phonetic transcriptions can be easily generated from the new Doc 9303 transliterated form (in exactly the same way as is done now from the Arabic name).

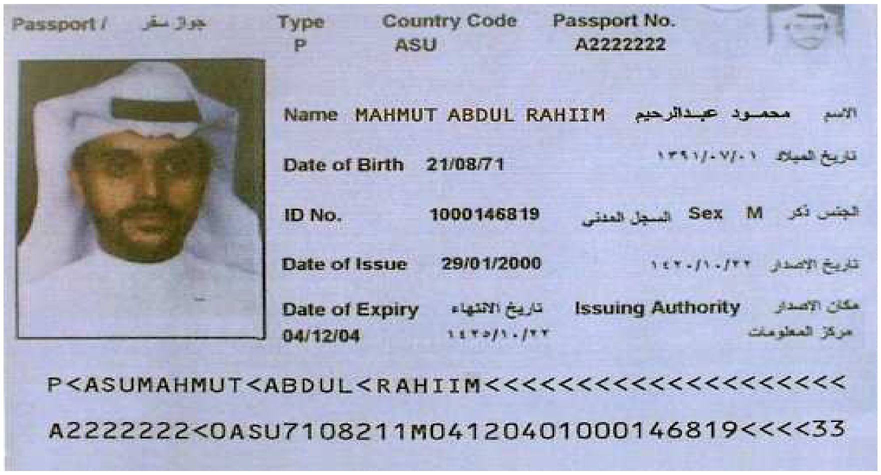
1.3 There are many existing transliteration standards for Arabic to Latin, however none of them only use the Latin characters ‘A’ to ‘Z’, and ‘<’, which is mandated by Doc 9303 for use in the MRZ. Therefore, much effort has been made to develop the transliteration method described hereunder to make it suitable for use in the Machine Readable Zone (MRZ).

2. TRANSLITERATION OF ARABIC CHARACTERS IN MRTDS

2.1 Doc 9303 states that if the name in the Visual Inspection Zone (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."

2.2 In the case of Arabic the transliteration is generally phonetic based and is termed a "transcription". The transcription is usually repeated in the (MRZ). This means that the advantages of the MRZ in terms of facilitation (fast and accurate reading of the passport data) accrue to countries that use the Latin script. For countries that use the Arabic script, the name is available in the phonetically transcribed Latin characters and <u>not</u> in the original Arabic script.



Passport / جواز سفر Type P Country Code ASU Passport No. A2222222

Name MAHMUT ABDUL RAHIIM محمـود عبدالرحيم الاسم

Date of Birth 21/08/71 ١٣٩١/٠٧/٠١ تاريخ الميلاد

ID No. 1000146819 السجل المدني Sex M الجنس ذكر

Date of Issue 29/01/2000 ١٤٢٠/١٠/٢٢ تاريخ الإصدار

Date of Expiry تاريخ الانتهاء Issuing Authority مكان الإصدار
04/12/04 ١٤٢٥/١٠/٢٢ مركز المعلومات

```
P<ASUMAHMUT<ABDUL<RAHIIM<<<<<<<<<<<<<<<<<<<<
A2222222<0ASU7108211M04120401000146819<<<<33
```

Figure 1: Phonetic transcription of Arabic name in both VIZ and NRZ

2.3 We can identify three disadvantages of the present transcription scheme:

a) The transcription is inaccurate, and this affects reliable identification;

b) The benefits of reading the MRZ are not available to countries that use the Arabic script; and

c) The present transcribed name is detrimental to the cultural heritage of the holder of the MRTD.

Example 1: Present variation in Transcriptions

The name in Arabic script:	محمود عبدالرحيم
A possible transcription into Latin characters:	Mahmut Abdul Rahiim
which is copied into the MRZ:	MAHMUT<ABDUL<RAHIIM
Number of variations[1] of "Mahmut":	>69 (Mahmoud, Mahmud, Mahmood, Mehmood...)
Number of variations of "Abdul Rahiim":	>144 (Abd-al-Rahim, Abdalraheem, Abd ar-Raheem, Abd al Raheem...)
Total number of possible variations in the name "Mahmut Abdul Rahiim":	>9,936
The name in Arabic recovered from the MRZ:	?

3. OCR-B AND ARABIC NAMES IN THE MRZ

3.1 While it can be claimed that the mandatory use of the Latin character based OCR-B in MRTDs is orientated towards a Western viewpoint, it must be remembered than when Doc 9303 was first published in 1980, OCR (OCR-A and OCR-B) was the only viable optical reading technology available. Paragraph 9.2.2 reads as follows:

> "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."

3.2 As well, the English language and Latin character based communications have become the standard for facilitation and security in the aviation sphere. Now, because of the installed infrastructure (machine readers, databases, etc) it is undesirable, and probably impossible, to change the MRZ, or adopt another technology to replace it.

3.3 The increasing use of ePassports means that it is possible to encode Arabic (and other scripts) in Unicode (ISO/IEC 10646) in the chip in DG11[2]. However not all countries will read the chips in ePassports, and of those that do, the chips in ePassports will not be read in every situation, so the MRZ remains the basic standard for machine reading. As well, the MRZ remains the back-up for situations where the chip fails to read (the MRZ is also necessary for gaining access to the chip using BAC).

3.4 ICAO has been aware of the importance of Arabic script for approximately for twenty four countries in North Africa, the Middle East, and South and Central Asia. These countries provide a sizeable fraction of the travelling public, both between these countries and to others. For a long time, the confusion and inaccuracy surrounding the phonetic transcription process has been known, because the Arabic characters:

a) represent some sounds not used in Latin-based languages;

b) represent some sounds that map to the same Latin character; and

c) vary in character sounds between the different Latin-based languages.

3.5 Phonetic transcription results in a large variety of Latin representations of the original Arabic name, sometimes in the order of thousands of variations for each name. This makes database matching problematical.

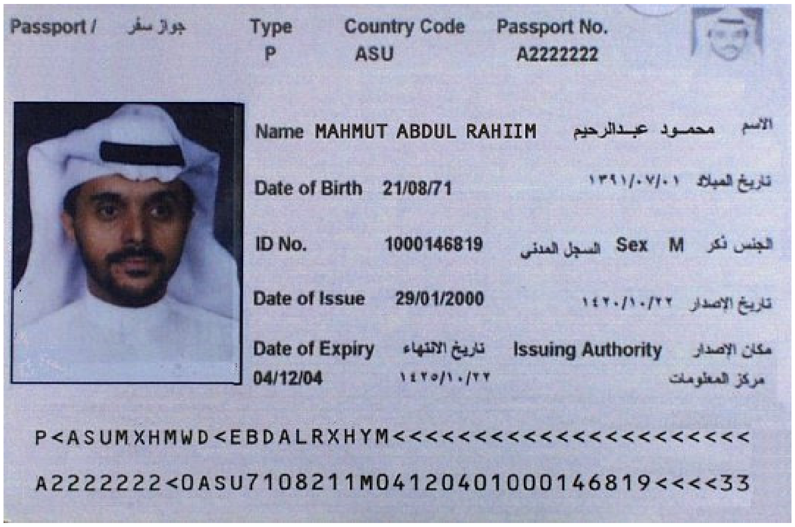
[1] From: http://www.kanji.org/cjk/arabic/araborth.htm

[2] Supplement Release 10: R10-p1_v2_sIII_0060

4. PROPOSED TRANSLITERATION SCHEME FOR THE MRZ

4.1 To achieve reliable identification, the original Arabic script name must be preserved. This also addresses the question of cultural heritage and enables the countries that use the Arabic script to receive the benefits of machine reading.

4.2 The Technical Report presented in TAG/MRTD-WP/17, 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 character so 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).



Passport / جواز سفر Type P Country Code ASU Passport No. A2222222

Name MAHMUT ABDUL RAHIIM محمـود عبـدالرحيم الاسم

Date of Birth 21/08/71 ١٣٩١/٠٧/٠١ تاريخ الميلاد

ID No. 1000146819 السجل المدني Sex M الجنس ذكر

Date of Issue 29/01/2000 ١٤٢٠/١٠/٢٢ تاريخ الإصدار

Date of Expiry تاريخ الانتهاء Issuing Authority مكان الإصدار
04/12/04 ١٤٢٥/١٠/٢٢ مركز المعلومات

```
P<ASUMXHMWD<EBDALRXHYM<<<<<<<<<<<<<<<<<<<<<<
A2222222<0ASU7108211M04120401000146819<<<<33
```

Figure 2: Transcription in the VIZ and transliteration in the MRZ

4.3 The result of using this look-up table process is a mapping of the original name in Arabic 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.

We can identify three advantages of the proposed transliteration scheme:

a) transliteration is accurate, and thus reliable identification is obtained;
b) benefits of reading the MRZ are available to countries that use the Arabic script; and
c) transliterated name preserves the cultural heritage of the holder of the MRTD.

Example 2: The proposed transliteration scheme

The name in Arabic script:	محمود عبدالرحيم
The new transliteration in the MRZ:	MXHMWD<EBDALRXHYM
Number of variations of "MXHMWD":	1
Number of variations of "EBDALRXHYM":	1
Total number of possible variations:	1
The name in Arabic recovered from the MRZ:	محمود عبدالرحيم

4.4 For countries that use the Arabic script, the name in the Arabic script is immediately available for display on computer screens and for database matching. This is possible using a simple computer program to perform the mapping, preferably to Unicode.

4.5 With respect to the ePassport, it is expected that the name in Arabic, using Unicode encoding, will be contained in the Data Group 11 (DG11) field of the chip. Therefore, there can be an immediate comparison for security purposes between the name in the MRZ and the name encoded in the chip. In fact, this becomes an important security factor in favour of the adoption of this scheme as otherwise the matching process cannot be easily done between the existing non-exact MRZ phonetic transcription and the name in Arabic script encoded in Unicode in DG11.

Example 3: Comparing DG11 with MRZ

Representation of the Arabic name in Unicode in DG11:	0645, 062D, 0645, 0648, 062F, <space>, 0621, 0628, 062F, 0627, 0644, 0631, 062D, 064A, 0645
Mapping the Unicode to Arabic characters:	0645=Meem (م), 062D=Hah (ح), 0645=Meem (م), 0648=Waw (و), 062F=Dal (د), <space>, 0621=Hamza (ء), 0628=Beh (ب), 062F=Dal (د), 0627=Alef (ا), 0644=Lam (ل), 0631=Reh (ر), 062D=Hah (ح), 064A=Yeh (ي), 0645=Meem (م)
The name in Arabic script from DG11:	محمود عبدالرحيم

The MRZ:	MXHMWD<EBDALRXHYM
Decomposing the MRZ:	M, XH, M, W, D, <space>, E, B, D, A, L, R, XH, Y, M
Mapping the MRZ transliteration to Unicode:	M=0645, XH=062D, M=0645, W=0648, D=062F, <space>, E=0621, B=0628, D=062F, A=0627, L=0644, R=0631, XH=062D, Y=064A, M=0645
Mapping the Unicode to Arabic characters:	0645=Meem (م), 062D=Hah (ح), 0645=Meem (م), 0648=Waw (و), 062F=Dal (د), <space>, 0621=Hamza (ء), 0628=Beh (ب), 062F=Dal (د), 0627=Alef (ا), 0644=Lam (ل), 0631=Reh (ر), 062D=Hah (ح), 064A=Yeh (ي), 0645=Meem (م)
The name in Arabic script from the MRZ:	محمود عبدالرحيم

4.6 While the ePassport offers the best way forward for encoding names in the Arabic script, the fact remains that standard MRPs will be in circulation for many years yet, perhaps as many as 20 or 30, and some countries may never adopt the ePassport.

4.7 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. For an example of computer generation of phonetic names, see the ARAN system (http://www.kanji.org/cjk/arabic/aran.htm).

4.8 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.

4.9 It must be noted that the Arabic name, as it appears in the MRZ in the proposed transliterated form, is not generally phonetic 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 machine the matching Arabic characters are substituted (as shown in example 3 above). 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. THE VISUAL INSPECTION ZONE (VIZ)

5.1 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. 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.

6. ADVANCED PASSENGER INFORMATION

6.1 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."

We propose that the IATA/CAWG statement be followed and the name, as read from the MRZ, should be used for API.

— END —