International Civil Aviation Organization

WORKING PAPER

TECHNICAL ADVISORY GROUP ON
MACHINE READABLE TRAVEL DOCUMENTS

Sixteenth Meeting

(Montreal, 26 to 28 September 2005)

Agenda Item 1:   Development of specifications for MRTDs
Agenda Item 1.3: Report of the Document Content and Format Working Group

TRANSLITERATION RULES FOR COUNTRIES
WITH NON-LATIN NATIONAL SCRIPT

(Presented by the Document Content and Format Working Group)

1.   **INTRODUCTION**

1.1   The Education and Promotion Working Group (EPWG) has traditionally been active in following up transliteration issues and developing guidance for countries who desire to transliterate their national script into scripts that conform to the rules of Doc 9303 for entries in the Visual Inspection Zone (VIZ) and the Machine Readable Zone (MRZ) of Machine Readable Travel Documents.

2.   **TERMS OF REFERENCE**

2.1   The EPWG has developed transliteration tables for the DCFWG that have been incorporated into Doc 9303 and that give useful and sufficient guidance for many countries, in particular for those with national scripts in the Latin and Cyrillic family. Beyond that, EPWG has worked on more general transliteration rules that are applicable in cases that are not covered by the above-mentioned tables. An Information Paper (IP/1) explaining these rules had been presented to the twelfth meeting of the TAG/MRTD in the year 2000, an updated version of which is attached to this Working Paper.

2.2   These general rules become more and more important as countries with scripts outside the Latin and Cyrillic family of scripts start to implement MRTDs. DCFWG has identified a need to give such general rules a greater publicity and has therefore reviewed the above-mentioned Information Paper.

It has been found fit for the purpose and, though not addressed to laymen, to be of great value for those in charge of the task to create transliteration rules in different countries, for the purpose of introducing MRTDs.

3.   **ACTION BY THE TAG/MRTD**

3.1   The TAG/MRTD is invited to:

    a) take note of the work of DCFWG with respect to general transliteration rules for countries;

    b) approve the publication of the Appendix to this paper on the ICAO web-site as a file that can be downloaded, in order to make it available to the general public; and

    c) recommend the use of rules as set down in the Appendix for the implementation of country specific transliteration schemes.

— — — — — — — —

Transliteration in ID-Documents

Non-technical abstract

This technical note explores the issue of transliteration, i.e. transferring a written item from one language and writing system to another. The issue is first considered broadly, then in the narrow context of international machine readable travel documents. One of the chief problems discussed in the paper is the difficulty in searching for a single transliteration scheme that would function between many languages. The paper concludes that this is not feasible, but does offer a set of proposed rules for choosing a specific scheme for transliteration between two specified languages.

The first major section of the paper, *names and language encoding*, explains *literation*, the process by which spoken language is encoded into writing systems. Three main processes of literation exist. Of these, the process known as *sound encoding* is explored in more depth, as it has most relevance here. Sound encoding makes use of alphabets, which are expanded by the use of *accents* and *diacritical marks*. As languages evolve and new languages are encoded into writing systems, increasing numbers of variations arise.

The second section, *methods of transliteration*, explores a range of methods which can be used for transliteration, i.e. expressing language in the coding scheme of another language. Options include one-for-one character substitution, substitution on the basis of meaning, and phonetically based methods.

The paper then discusses *phenomena in transliteration* in a third section. The problems that arise in transliteration are summarised as follows:
- the sound value of a character is usually dependent on its context;
- some sounds in one alphabet do not have an equivalent in the other;
- the same letter(s) may have a different sound value in different languages;
- more than one letter may be needed to define a sound;
- it may be difficult to ignore traditional transcription schemes.

All of these problems point to the fact that letter-by-letter transcription schemes are very limited, and it is generally better to use a sound-based, phonetic scheme. They also indicate the complexity of the transliteration task and the importance of attending to the specifics of the two languages involved. While reciprocity is an ideal (i.e. using the second transliteration scheme, reversal of the first transliteration, would give the exact original piece of language), it is seldom achievable. Likewise, universal transliteration schemes are not considered to be a realistic aim.

The fourth section of the paper focuses on the issue of setting *objectives of transliteration in ID-documents*. In this context the sound encoding of names is primary; meaning has little relevance. The limited character set allowable in the MRZ (machine-readable zone) will pose some challenges, and data records will need to ensure that personal data is linked as it passes through the stages of transliteration.

A fifth section considers *suggestions for transliteration rules*. This section gives flow diagrams setting out the logical process of transliteration. In general, two transcription schemes are needed to interface between any two languages and their respective writing systems. In addition, separate transcription schemes may be used for the VIZ and the MRZ. The relationships between the various transcription schemes are explained in some detail.

This section of the paper also contains a set of proposed rules to be used when creating or selecting a scheme for transliteration between two languages. These rules are not summarised here, as they are highly specific, so should be consulted in the original document.

Summary

Problems in writing systems and their mutual transliteration are examined and analysed in view of the needs of transliteration in international machine readable travel documents. A set of rules is proposed that could be used as a guideline for territorial or international authorities when choosing or developing transliteration methods. This set of rules is not prescriptive with respect to specific methods to be used for transliteration from one language to another; they rather aim to outline best practice in executing such implementations. The additional problem of truncation arising from space limitations in the documents is not dealt with in this paper.

Preface

The actual set of standards that reflects the original ICAO document 9303 and its subsequent evolution is highly prescriptive with respect to the representation of visual and machine readable data. The Latin alphabet, the use of which is the essence of these standards, fails - with or without its various extensions for the use in special language environments - to meet the tremendous variety of expectations and requirements expressed by users in a world-wide multicultural environment. On the other hand, limitation to a character set designed for a maximum of interoperability for human as well as for machine-based interpretation is the only possible choice in order to meet the goal of ensuring world-wide acceptance of a single-document-based identification scheme. The group of experts within ISO which has been monitoring and supporting ICAO's standardisation efforts from

2

the outset has discussed the pros and cons of alternatives on many occasions and has again and again come back to the conclusion that the restrictions given in the actual standards are essentially valid.

The purpose of this paper is to explain the role and objective of transliteration in this context and to provide guidance for those who are responsible in their respective countries and organisations for specific document systems of the ICAO doc. 9303 type. This paper is not prescriptive but should help those responsible to define individual transcription rules in a smooth and appropriate way. The approach is to give a basic understanding of the problems of transliteration, covering all relevant aspects including technical, historical, legal and cultural aspects, and to justify the proposed guidelines and rules from this background.

Names and language encoding

Literation

The use of language as a means of communication is one of the basic human natural skills. Writing as a means of conservation of language constructs for repeated use and verification is a consequence of cultural skills and conventions that have been developed for only a relatively restricted set of languages. Out of an estimated number of 5000 existing languages only about thirteen percent have ever been systematically encoded into a consistent set of rules for graphical representation. Nevertheless, virtually every individual on this globe will, during his life-time, at least have his name and the name of his place of origin and residence encoded in one of the existing sets of "rules for graphical representation for language constructs". Therefore, before we talk about transliteration we will have to talk about literation in the first place.

Those languages that are consistently related to at least one such set of rules may be called the literated languages. The respective set of rules may be called a *writing system*. People coming from or living in a cultural environment that makes use of a non-literate language, will unavoidably have to rely on an existing writing system in order to encode their names. This process of borrowing on another language's writing system may be called *paraphrasing*. In many countries there is a well established writing system that provides a defined character set, rules of orthography and typesetting conventions, so that paraphrasing seems to be a simple task. Nevertheless, problems and disputes may arise out of this task, especially if the original non-literate language contains sound constructs that are not well represented in the language used for paraphrasing, or if, out of historical, political or cultural background, different paraphrasing options exist.

Language encoding systems

The language systems in use today for literation and paraphrasing may be distinguished in the three categories of

3

➢ *term encoding*: The graphical elements of the writing system represent a specific meaning, irrespective of the details of pronunciation. The most important example is the family of Chinese characters.

➢ *syllable encoding*: Typical sound sequences of a specific language are represented in distinct graphical elements. The most popular modern examples are the Korean and Japanese writing systems.

➢ *sound encoding*. At first sight the most precise and elementary description of human speech, using one character per basic sound element, sound encoding is one of the prime sources of "wrong pronunciation" because of the big differences in language-specific intonation of more or less similar sounds and, as a consequence, the different ways that people with different literation backgrounds interpret the written character sequence.

In addition to their specific graphical elements, called *character sets*, writing systems comprise language-specific schemes of spelling, pronunciation, accentuation and punctuation (altogether referred to as *orthography*) and more or less well defined types of *scripts*, which are above all adaptations of the graphical appearance of characters according to different needs, but may also comprise modifications in orthography and other rules of usage.

Sound encoding

As the present context will be primarily focussed on sound encoding schemes, it is not necessary here to go into details of term encoding and syllable encoding schemes. When looking at sound encoding schemes it is worthwhile to note the following details:

Sound encoding schemes are, in principle, not limited to use in a single language. On the contrary, the world-wide predominance of sound encoding schemes today is to a large extent a consequence of their ease of adaptation to the needs of various languages. A second inherent advantage of sound encoding schemes is the simplicity of the character set. Sound encoding character sets are called *alphabets*. Most existing alphabets comprise between twenty and thirty characters. This facilitates the achievement of high levels of literacy in a given population as well as the implementation of simple mechanical writing tools (typewriters, keyboards). Last but not least this inherent simplicity of sound encoding justifies its use in the environment considered here, i.e. in a universal encoding scheme for personal identification documents. A burden that has been added to sound encoding schemes in the course of their history, in the interests of making them useable in multicultural environments, is the use of *accents* and *diacritical marks*. Exaggerated use of such "modifiers" to the meaning of characters in some languages complicates the use of the alphabet to an extent that gives reason to wonder whether it would not have been easier to use other means of encoding than a sound-oriented scheme. On the other hand, in most cases careful analysis shows that simple sound encoding schemes without modifiers could have done the job for any specific language perfectly well; however perhaps those who

4

created the actual schemes were biased in their approach by their own cultural environment or were too much focussed on a perfect letter by letter representation of sounds without having recourse to clever pronunciation rules. A vivid example of such complicated schemes, that may multiply the size of the effective character set with respect to the original alphabet, is the Latin encoding of the Vietnamese idiom. Unfortunately it is, in most cases, unthinkable to change an existing scheme, because of the large background of existing literature and other accumulated data records.

In the classification of alphabets several levels of parenthood should be taken in consideration. The following overview lists different aspects contributing to differences and similarities:

Creative approaches

The majority of alphabets in use today are derived from Latin, Greek or Arabic alphabets. In many cases, however, the creators of secondary alphabets such as that which is in use in Russia today, made use of the pool of existing letters in other alphabets as raw material. They attached to them new interpretations of their sound value and felt free to add new characters and to omit others. The consequence is that identical characters may express completely different sounds in different languages and vice versa. For example, the Latin characters B, C, H, N, P, R, X appear in the Russian alphabet with a completely different sound value, whereas the original sounds of these letters either do not exist in Russian or are represented by symbols not existing in the Latin alphabet, such as П. Latin letters like F, G, J, S, L are not used in the Russian alphabet. (Similar correspondences and differences may be found in comparing the original Latin and Greek alphabets, as they were derived from earlier alphabets in a similarly creative way.

Examples of historical evolution

The original Latin alphabet is the ancestor of most "western" alphabets of today. The original Latin alphabet comprised no more than 20 characters and had no accents or diacritical marks. In two initial evolutionary steps five more characters were introduced, the first by cultural influences from Greece, the others by gothic invaders when grafting the needs of their own native languages onto the Latin coding scheme. Accents became popular in the great times of medieval hand writing art, at first mainly to reduce the length of words and thus to speed up the tedious work of writing. Eventually accents became an integral part of the orthographies of several languages that were created during that time. The same is true for the typical diacritical marks specific to Nordic and some Slavic languages.

The transliteration approach

Some cultures have, at some point in their history, adopted an alphabet as used in another suitable language and have in this way either reached the level of literacy for the first time or abandoned earlier coding schemes. An example is Turkey, which introduced the Latin

alphabet at the beginning of the twentieth century using a transliteration scheme derived from the German language and at the same time abandoned an alphabet derived from the Arabic. Such close transliterations have a tendency to introduce complicated accent schemes that would not be necessary if the creative approach had been chosen, even without the creation of new characters.

Methods of transliteration

Characters and rules

Similarities between different alphabets are mostly the consequence of earlier transliteration schemes. In general, transliterations may be considered as more or less arbitrary, more or less efficient, and more or less successful attempts of expressing a language within the framework of a pre-existing coding scheme. In order to become a consistent writing system, the transliteration must become independent from its original framework. As mentioned above, a writing system is not yet completed by defining a character set and allotting a sound value to each character. Rather, a complete set of rules is required to eliminate ambiguity and inconsistency in spelling and accentuation. In general, the rules cannot be derived from those used in the original language context of the writing system, because those were specific to the original language. Each language has its peculiar nature. Consequently, it is almost always necessary to create specific orthographic rules when creating a transliteration scheme. To understand the options that can be taken for transliteration, it may be useful to have a closer look at the mechanisms involved in the coding of language into writing.

Principles of language encoding

To keep this discussion (which is not central to the topic of the paper) as simple as possible we limit ourselves to character based sound encoding schemes. Fig 1 below illustrates the typical sequence of actions involved in the generation of a written representation of language. Coding is, in general, not only a representation of sound elements by symbols taken from a character set. Useful coding depends on the meaning of what is said and therefore involves the application of more or less complicated rules that account for the context in which a specific sound sequence is pronounced. Real encoding will very likely have to account for accents and diacritical marks as well and may have to take into consideration the specific graphical subtleties of the character set such as ligatures, ornamental initials etc and is therefore even more complex than the schematic approach shown in Fig 1. The complete logical set is called a phonetic system (see also the section "Suggestions for transliteration rules" below).

```
                ┌──────────────┐          ┌──────────────┐
                │ Analysis of  │          │  Character   │
                │   context    │          │     set      │
                └──────────────┘          └──────────────┘
                       │                         │
┌──────────┐    ┌──────────────┐   ┌──────────────┐   ┌──────────┐
│ Phonetic │    │ Extraction of│   │  Coding of   │   │          │
│  sample  │───▶│   meaning    │──▶│  character   │──▶│ Coded text│
│          │    │              │   │  sequence    │   │          │
└──────────┘    └──────────────┘   └──────────────┘   └──────────┘
                       │                  │
                ┌──────────────┐   ┌──────────────┐
                │ Analysis of  │   │   Set of     │
                │    sound     │   │ orthographic │
                │   elements   │   │    rules     │
                └──────────────┘   └──────────────┘
```
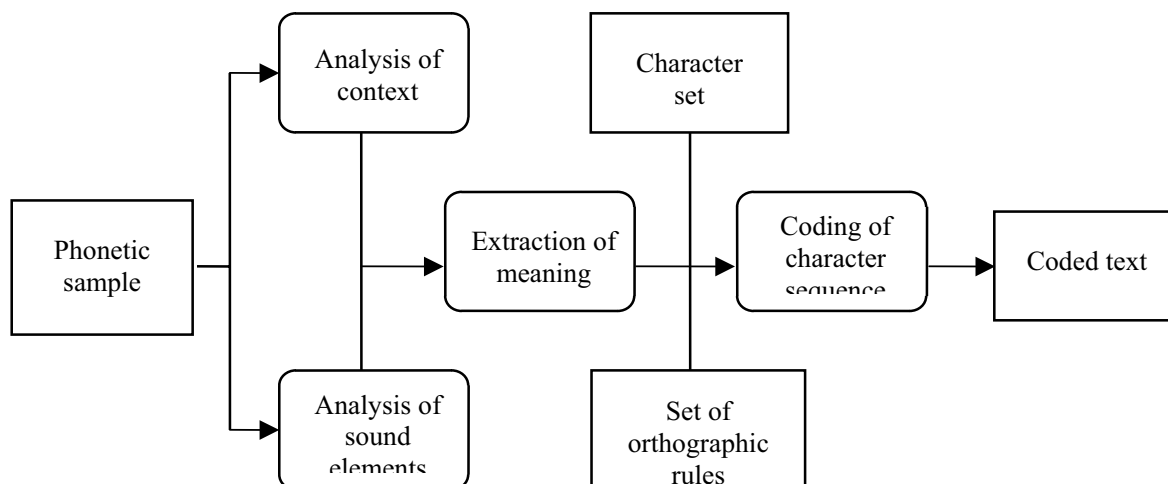
Fig 1: Flow chart of writing code generation.

If we consider *transliteration* with diagram of Fig 1 in mind, it becomes apparent that the bridge from one character representation to the other may be constructed at any point of the flow chart, depending on how translation rules are applied. One has to imagine two diagrams like Fig 1 and then to look for possible means of "borrowing" from preexisting patterns of one diagram for use in the other. For transcribing a coded text, originating from language L1 and represented in a character set C1, into a secondary text using a different character set C2, several of the following options may be viable.

Character replacement

If a one to one relationship exists between the two character sets, i.e. a list can be established which relates each single character of set C1 to a different single character of set C2, it may be legitimate to ignore all rules that normally apply to the use of C2 in its natural language environment L2 and to simply replace the characters according to the relationship established in this list. Once applied, the transcription may be reversed and the original representation in C1 may be re-established by applying the list in the inverse direction. Such literal transcription schemes, however, will be limited to alphabets where the character count of C2 is equal to or higher than that of C1. Other limitations and drawbacks will become apparent in the next section (Phenomena in transliteration).

Meaning based replacement

Such a scheme would go to the other extreme and simply ignore the actual C1 representation of the text and would create its own orthographic rules that are considered adequate to cope with the limitations of C2. Such an approach seems to be the most appropriate whenever

7

the nature of the character sets C1 and C2 is too different to be used for a one by one replacement. If desired, the new rules may be constructed to show some similarity to rules already existing in the context of C2 with respect to other languages. For example, it was an arbitrary decision of the creators of the new Turkish alphabet to consider German sound coding rules when adapting the Latin alphabet and creating a set of orthographic rules for the representation of Turkish text in that alphabet. Transliteration is, in this case, very much the same as the paraphrasing mentioned earlier in this paper, without paying too much attention to the original Arabic spelling. It is not possible to reconvert the Latin expression into the original Arabic spelling without thorough knowledge of both literation rules.

Phonetic Imitation

Sound based coding rules are, in general, only one part of the orthographic system of rules. They may nevertheless be considered satisfactory in cases where the actual meaning of a text element seems to be of little relevance, as is the case with lists of names. As the vocal interpretation of characters is rarely unambiguous, it is necessary to establish clear pronunciation rules in the actual language context L1 with respect to its representation by a character set C2. Of course it may nevertheless be useful to adopt pre-existing pronunciation rules taken from language L2 in order to minimise necessary adaptations.

Phenomena in transliteration

Historical inconsistencies

Language is a "moving target". Rules, once set in a rigid manner, will undergo historical development. When two languages rely on the same set of characters, diverging developments may lead to inconsistencies, even if the original approach was quite adequate. Take as an example the word "Europe". It came to exist as a Latin transcription of a corresponding Greek name "ΕΥΡΩΠΗ". Already the Romans took some freedom in interpreting the original sequence of Greek letters by systematically replacing the Y with U in double vowels. If one looks at the pronunciation of this old word in different languages of the geographic area described by it, one will find a tremendous variety of sounds. Comparing transliterations that try to imitate the English way of spelling, we find today in Greece "Avropi", in Italy and Spain "Aooropa", in France "Erhopp", in Germany "Oyropa", and finally the English "Yourop". However because of respect for tradition, nobody would consider making such transcriptions. It is interesting enough that respect for tradition does not stop modern Greek transliterators from transforming the famous French philosopher Decartes into "ΔΕΚΑΡΤ", or re-transcribed "Dekart", for the simple reason that at some point the Greeks decided to opt for a phonetic transcription of French names in view of the fundamental differences between the Greek and French alphabets and the difficulty of finding an adequate universal literal transcription. Still, when transliterating a French word sequence containing "Europe", the Greeks certainly would not write "ΕΡΟΠ",

8

which would correspond to the official phonetic transcription scheme, but would not hesitate to put the original "ΕΥΡΩΠΗ". This is to shed some light on the fact that any formal transliteration scheme will permit ambiguities where historical context is relevant.

Using composed letters

To mention only one example, the two letters "sh" are usually not meant to be pronounced separately in English but will form one specific sound that is not represented by a corresponding single letter in the English alphabet. When transcribing an English expression containing *sh* into a language that has a specific *sh*-sound available, one would probably expect that the specific sound will replace the two letters sh. In the same way, knowing that Greek "ΟΥ" is pronounced "u", one would probably suggest to use "u" instead of "oy", in the corresponding English transcription. This is to shed light on the fact that in most cases transliteration requires more than just letter-by-letter replacement. Rather, rules need to be established that take care of the specific relationship and difference that exists between two writing systems.

Reciprocity

Following the rules cited above in "Examples of historical evolution", when you transcribe "Veranda" into Greek this will give "ΒΕΡΑΝΤΑ". Transcribing it back, it may well become "Beranta", if no special care is taken with respect to the specific sound value of Greek characters depending on their position in a word. This again should be reconverted back into "ΜΠΕΡΑΝΔΤΑ" if Greek rules are correctly applied. The current transcription schemes do not guarantee reciprocity. Special care is necessary if one wants to ensure that a transcribed text will revert to its original when the transcription is reversed.

General situation

These few examples should be sufficient to understand the general situation. Transliteration is a very complicated subject even in the simple case of two closely related sets of alphabet rules. The list of problems may be summarised in the following general statements:

➢ The sound value of a character may be - and usually is - context dependent
➢ Some sounds in one alphabet do not have an equivalent in the other
➢ The same letters may have different sound value in different languages
➢ More than one letter may be needed to define a sound
➢ It may be difficult to ignore historical transcription schemes

9

Conclusion

As a consequence of the problems described above it may be useful not to choose a letter-by-letter transcription but a sound-based, i.e. phonetic one. Letter-by-letter transcriptions have a very limited scope.

Any attempt to define a single transcription scheme for alphabet pairs which relate to more than one pair of languages will seldom give reasonable results.

Transcription is a matter with many factors involved. If a certain level of general acceptability between any two of the phonetic systems is to be achieved, the transcription must attend to all particularities of the phonetic systems involved. And this does not seem to be possible using a universal letter-by-letter transcription scheme.

Objectives of transliteration in ID-documents

When setting the more limited objectives for transliteration in ID-documents the following background must be kept in mind:

Reproducing names by sound

Transliteration rules may be entirely limited to names, either names of persons or names of places. In most cases there is no specific meaning associated with names. The aim should be to achieve a transcription that will give a hint of how to pronounce the name in its original language context. As a consequence, the transcription should refer to the pronunciation rules of a specific language currently in general international use. Given the language practices of ICAO and ISO this will mean that one of the languages English, French or Spanish should be used for this reference. Any historical background of names should, in this context, be considered as less important and therefore be ignored whenever it is in conflict with the objective of correct pronunciation.

Making best use of the restricted MRZ character set

For good reasons neither accents nor diacritical marks are allowed in the Machine Readable Zone (MRZ). The only one of the three above-mentioned languages that is mostly free of such character set extensions is English, though it is also good practice in French to omit accents and diacritical marks in texts written in capital letters. There might be good reasons for the issuing authority to use two sets of transcription rules: the one for best guidance in pronunciation within the VIZ (Visual Inspection Zone), the other for best use of the limited character set for machine based search criteria within the MRZ. It could be confusing, however, if entirely different approaches were chosen for the two sets of rules. The best approach will probably focus on a fine-tuned, even complicated set of rules for the

generation of VIZ data, with a simple, even automatic derivation of MRZ data from the data represented in the VIZ.

Facilitation of electronic data base search

In many cases users will want to rely on MRZ data for purposes of computer-based storage and retrieval of information. In order to avoid ambiguity it is desirable that personal records, beginning with those in documents in the original language and ending with the entries in the MRZ, all have a transparent and unique correlation. Ideally, this correlation will be reciprocal. However, reciprocity is difficult to achieve and in most cases is not really necessary. Therefore, the set of rules given in the next section includes the objective of reciprocity; however this must generally be disregarded because of lack of feasibility. Rather, the focus is on establishing transparency and avoiding ambiguity, at least in a one-way transliteration scheme.

Suggestions for transliteration rules

A formal approach

Most of the suggestions that can be derived from the initial analysis are already implicit in the concluding remarks given in the last section. In order to reach conclusions that can be used universally, this section will attempt to establish a set of more formalised rules.

Let us look at two phonetic systems Ps_1 and Ps_2. As a general rule, it will be necessary to develop two specific transcription schemes T_{12} and T_{21} in order to interface between these two systems; one that transcribes from Ps_1 to Ps_2, one that transcribes back from Ps_2 to Ps_1. When transliterated, a specific text expressed in Ps_1 will have a graphical representation R_1, expressed in Ps_2 with a graphical representation R_2. In a formal symbolic notation the act of transcription may be expressed as:
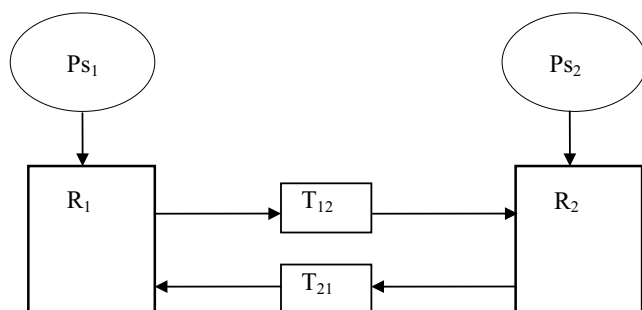
$$T_{12}(R_1) = R_2; \; T_{21}(R_2) = R_1.$$

Fig 2: Representations and transcriptions.

Ideal (but mostly unachievable) transcription schemes should be reciprocal, i.e. if a pair of transcriptions T_{12} and T_{21} is applied in sequence, the result should be the original one:

$$T_{21}(T_{12}(R_1)) = R_1;$$
$$T_{12}(T_{21}(R_2)) = R_2.$$

Again , given the fact that existing phonetic systems are always very specifically adapted to the needs of the language they are designed for, it is, as a general rule, virtually impossible to fulfil these requirements when many languages are involved. Therefore, and this is important to keep in mind, most existing transliteration schemes (T_{12}, T_{21}) do not guarantee (in fact it is highly unlikely) that consecutive application of T_{12} and T_{21} will lead back to the original representation as may be implied when looking at Fig 2. What will usually happen is an accumulation of oddities as described in the example of the "veranda" in the section "Reciprocity" above. In order to properly understand the arguments following below, it is important to see any representation R_x as a set of abstract entities. Transcribing members of a set R_1 to R_2 using T_{12} will result in a subset of R_2 which, when re-transcribed to R_1 using T_{21} will result in members of R_1 distinct from the original ones unless the specific design of T_{12} and T_{21} really provides reciprocity. Irrespective of those difficulties we will claim the goal of reciprocity in Rule 1 below, but will at the same time keep in mind that this is rarely achievable.

With respect to the requirements of a unified international transcription scheme as intended for the use in the MRZ and in the VIZ of an MRTD, we may look at the following sub-schemes for the transcriptions:

T_{OF} is the scheme that is generally used in the country issuing the document to create the *formal* representation R_F of inscriptions of the type later to appear on the VIZ in its own public registers. This may or may not be identical with the *original* representation R_O that is used by the individual to write down the same data and may also imply certain rules of transcription with respect to special characters that are part of the cultural heritage of the

12

country or the individuals living in it, but are not adopted for use in administration. This scheme T_{OF} is therefore not necessarily a transcription in the sense that it mediates between one consistent phonetic system to another. It might as well be a phonetic system of its own.

T_{FV} is the scheme that mediates between the written representation R_F given by T_{OF} to the one that is used in the *VIZ*, i.e. R_V.

T_{FM} in turn is the scheme that mediates between R_F and R_M, the latter being the representation used in the *MRZ*.

T_{FC} is a scheme that mediates between R_F and another scheme R_C that corresponds to an existing formal representation used in another *country*.

Fig 3 below displays the relationship between these representations and transcriptions. The 15 Rules given further below may be better understood by referring to this diagram.
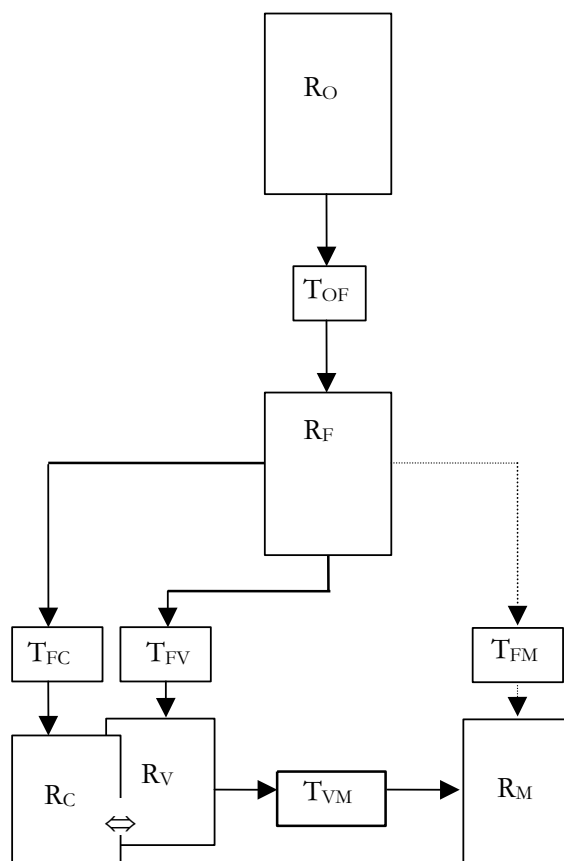
Fig 3: Representations and transcriptions

In particular, attention should be paid to the fact that the transcription T_{FM} is only inserted with dashed arrows. This is to indicate that such a direct transcription should not be explicitly used. The correct process, according to the Rules, would be to use T_{FV} and T_{VM} in sequence.

The representations R_C and R_V will be identical if the Rules below are fulfilled, and are therefore shown as overlapping and linked by an arrow in Fig 3.

Resulting rules

Depending on the language, two or all of the representations R_F, R_V and R_M may be identical. However, this is only valid for a minority of countries. The vast majority will be confronted with considerable difficulties to find satisfactory transcription schemes that fulfil the objectives as set out above. The following rules may be established as a useful means of guidance for the selection and, if necessary, creation of transcription schemes mediating between the three basic representations R_F, R_V and R_M.

General

Rule G1)   All transcription schemes should be designed such that a simple scheme T_{VM} exists that fulfils the following relationship: $R_M=T_{FM}(R_F)=T_{VM}(T_{FV}(R_F))$.

Visual zone

Rule V1)   The transcription scheme T_{FV} should be identical to an existing national standard T_{FC} of the issuing country, if such a standard is available.

Rule V2)   The transcription scheme T_{FV} should be selected in such a way that the resulting representation $R_V=T_{FV}(R_F)$ only uses characters from the character set defined in the International Standard ISO 1073/2.

Rule V3)   If Rule V2 is not consistent with any national standard according to Rule V1, the character set used in R_V may be extended by introducing additional accents and diacritical marks.

Rule V4)   If no standard exists that fulfils Rule V1 together with Rules V2 or V3, another - non national - standard T_{FC} may be selected, provided that it is compatible with the restrictions imposed on the character set as set down in doc 9303 for the VIZ, with modifications of the kind mentioned in Rule V3 if necessary.

Rule V5)   If available, a standard T_{FC} according to Rule V1 or Rule V4 should be selected that can mediate between R_F and one of the official phonetic schemes R_C used

in the context of one of the following languages: English, French, Spanish.

Rule V6)   The transcription standard chosen according to Rule V5 should either be an international standard or a national standard of a suitable country.

Rule V7)   If the above-mentioned Rules are not able to provide a satisfying transcription scheme T_{FV}, it is strongly recommended to take steps to create a new standard or at least ad hoc transcription rules fulfilling the above requirements.

Rule V8)   If there is freedom of choice while respecting the above Rules, the transcription T_{FV} should be designed in such a way that a reciprocal transcription T_{VF} arises, following the equation: $R_F=T_{FV}(T_{VF}(R_F))$.

Rule V9)   If Rule V8 can not be fulfilled and there is freedom of choice while respecting the other Rules, the transcription scheme T_{FV} should be chosen in such a way that there is no ambiguity in creating R_V from R_F, i.e. there are never two different representations R_V that are both consistent with a single representation R_F.

Rule V10)   If there is freedom of choice while respecting the above Rules, the transcription T_{FV} should be chosen such that the phonetic value of the resulting representation $R_V=T_{FV}(R_F)$ is similar to the original phonetic value if the pronunciation rules of one of the following languages are applied: English, French, Spanish.

Rule V11)   The captions used for the designation of fields in the VIZ should be given in the language that is applicable according to Rule V10.

Machine readable zone

Rule M1)   The transcription scheme T_{VM} according to Rule G1 should be chosen in such a way that there is no ambiguity in creating R_M from R_V, i.e. there are never two different representations R_M that are both consistent with a single representation R_V.

Rule M2)   If there is freedom of choice while respecting the above Rules, the transcription scheme T_{VM} should be chosen in such a way that it follows existing standards in use with the language chosen according to Rule V10.

Rule M3)   The representation R_M must at any rate be compatible with ICAO doc 9303. No exception in the choice of characters can be tolerated without compromising the whole approach of global interoperability.

15

Conclusion

The analysis of the structure of existing writing systems that is given in this technical note suggests not to impose one universal transcription scheme for the use by all countries, but to leave it to each individual country to find its own appropriate solution. Nonetheless, a set of general rules may be established to facilitates this task. What this information paper does not attempt to do is to provide a text that is easy to read by laypeople. It may be advisable for decision makers to systematically rely on expert advice in these matters.

Simplified Flowchart for Transcription Scheme Design

Is transcription
necessary?

Does a national
Transcription
Scheme exist?

yes   no

yes

no

Does an
international TS
exist?

Existing Scheme
applicable in the
context E, F, S?

yes

no

yes

no

Design a new
Transliteration
Scheme

Any non E, F, S
Transcription Scheme
applicable?

Select
the most
appropriate
TS

no   yes

End

17