Thesis

Gabe DeFreitas

October 22, 2018

Contents

0.1 Introduction

> You see, to some people in the world, your name is everything. If I say my name to an elder Hawaiian (kupuna), they know everything about my husband's family going back many generations...just from the name.
>
> (Janice "Lokelani" Keihanaikukauakahihuliheekahaunaele)

0.1.1 Names

A name is a fundamental identifier of a person *qua* individual; as such, it seen as a highly personal issue, and the choice of name by the child's parents may carry cultural, genealogical, religious, linguistic information.

0.1.2 Computers and Writing Systems

The "digital age" has changed writing from a free and individualized practice to a one based in a discrete and logical structure consisting of discrete glyphs. Since most early development of computers took place in the United States, English gained a natural ascendancy over other languages in the field of digital communication. English, perhaps as a coincidence, is also one of the easiest languages to represent in code, requiring at the bare minimum just the 26 non-accented characters of the English alphabet, perhaps with some punctuation and numbers. As computer technology spread far beyond the labs of Silicon Valley, there were some associated growing pains with the need to encode a vast array of linguistic forms, including diacritical marks, right-to-left scripts, and pictographic writing systems.

0.2 Very Long Names

0.3 Birth Certificates

American law generally holds the naming of children to be the right and responsibility of parents, without shutting the door on regulating edge cases, like "Ghoul Nipple", "Legend Belch", "Brfxxccxxmnpccccllmmnprxvclmnckssqlbb11116", and "" [**larson11**]. In many American states, however, this right is abridged with reference to diacritical marks above letters, hardly an edge case in many languages around the world. **larson11** investigates this in his study of American naming law, finding states with such rules to include California, Massachusetts, New Hampshire, and Kansas:

0.3.1 California

Guidelines provided by the California Office of Vital Records instruct county clerks that baby names may contain only "the 26 alphabetical characters of the English language with appropriate punctuation if necessary" and that "no pictographs, ideograms, diacritical marks (including 'é,' 'ñ,' and 'ç') are allowed" [**larson11**].

The handbook cites Proposition 63, the 1986 ballot referendum in which Californian voters declared English the state's official language, as legal justification. Larson points out, however, that the names of two California state parks, Año Nuevo State Park and Montaña de Oro State Park, manage to contain such characters. Moreover, the City of San José, California includes the acute accented é in its official name, and its Style Guide even includes instructions on how to produce it electronically: "To create an accented é, hold down the alt key and type "0233", on the numeric key pad." California's Department of Public Health likely disobeys the city's guidelines in birth certificates, though this needs to be verified.

Proposition 63

A ballot initiative in 1986 declared English the official state langauge of California, supported by 74% of the voting electorate.

2014-AB-2528

A 2014 bill in the California State Assembly sponsored by AM Nancy Skinner (AB-2528) sought to rectify the state's processing of birth certificates and driver's licenses by allowing diacritical marks in names. The bill "required the

State Registrar to ensure that diacritical marks on English letters are properly recorded on birth certificates, death certificates, certificates of fetal death, and marriage licenses, including, but not limited to, accents, tildes, graves, umlauts, and cedillas" [AB-2528]. This bill stalled in the Appropriations Committee when state agencies predicted multi-million dollar price tags relating to IT upgrades, noting that the DMV's software could not "even accept lower-case letters". For this same reason the bill was opposed by the County Recorder's Association of California.

2017-AB-82

In 2017, California AM Jose Medina revived the issue with AB-82, which ultimately passed both houses of the legislature before being vetoed by Governor Jerry Brown. Unlike the 2014 bill, this edition did not affect the issuance of driver's licenses, only birth certificates. Passing through many more stages of the legislative process, the committee hearings gathered more detailed estimates for the cost of IT upgrades than they had in 2014:

- $230,000 for IT upgrades at Department of Public Health
- $2 million per year for Department of Public Health to correct existing records
- Loss of revenue of $450,000 per year to Department of Public Health since they would not be able to electronically transmit names to SSA (at $3 per name) containing diacritics
- Up to $12 million for local governments to upgrade their systems
- $1–3 million in upgrades to Department of Health Care Services
- Unknown administrative costs to Department of Social Services

The sticking point for Governor Brown was compatibility with federal databases, which do not accept diacritics. In his veto message, he argued that the risks to vital records outweighed the benefits of cultural openness:

"Mandating the use of diacritical marks on certain state and local vital records without a corresponding requirement for all state and federal government records is a difficult and expensive proposition. This bill would create inconsistencies in vital records and require significant state funds to replace or modify existing registration systems."

The committee findings make clear that the state would incur nontrivial costs to update the name registration systems. Little discussion is included of the possible creative solutions to the problem. Even assuming that government systems cannot be made to support the full UTF-8 standard, there are ways of representing information using ASCII. For example, we will see later that the international specification for machine-readable passports has a variety of control

sequences for representing subtle distinctions in the Latin, Cyrillic, and Arabic alphabets using only the 26 plain characters of the English alphabet. The original form can be recovered nearly losslessly using the transliteration table.

0.3.2 Massachusetts

In Massachusetts, the "characters have to be on the standard american [sic] keyboard. So dashes and apostrophes are fine, but not accent marks and the such". [Larson11]

0.3.3 New Hampshire

"All special characters other than an apostrophe or dash" are prohibited. [Larson11]

0.3.4 Kansas

Restrictions are similar to those in Massachusetts. [Larson11]

0.4 Passports

The protocol governing machine-readable travel documents (MRTD) is Document 9309, issued by the International Civil Aviation Organization [**ICAO9309**]. These standards define the common form that all passports must take to ensure interoperability. Since all states must operate on a shared standard, the diplomatic community has forged a compromise between cultural diversity and international security; the 9309 standard provides sufficient flexibility to accommodate the diverse languages and scripts used worldwide.

A 9309-compliant MRTD data page is divided into two sections: the Visual Inspection Zone (VIZ) and the Machine Readable Zone (MRZ).

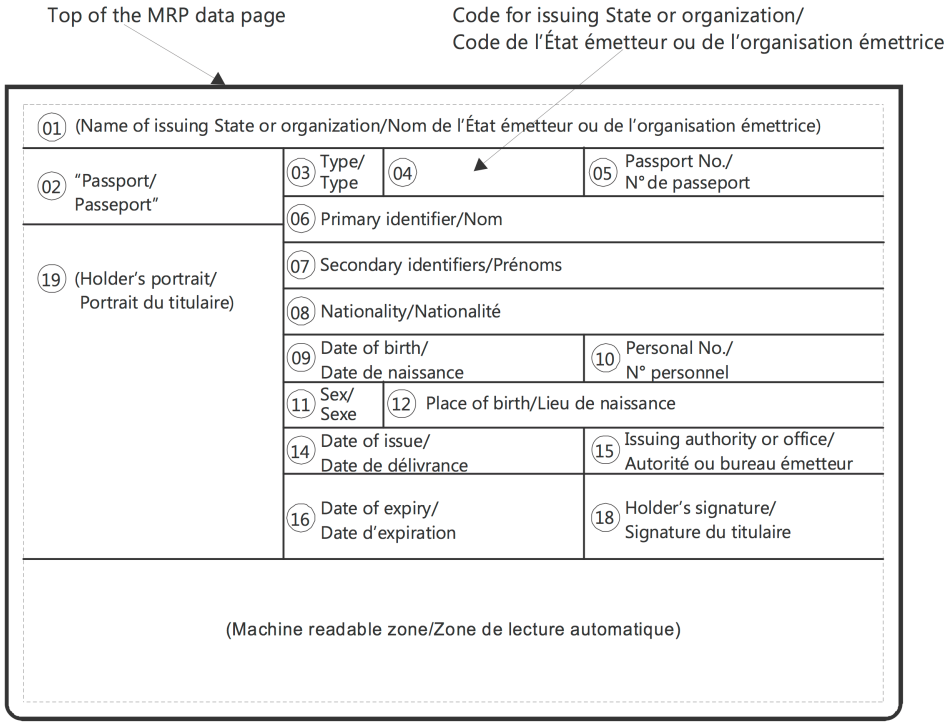
0.4.1 Visual Inspection Zone

The Visual Inspection Zone, consisting of the top two-thirds of the passport's data page is designed for inspection by border officials at the point of entry. States may populate the required VIZ fields in their official language provided a translation is provided into English, Spanish, or French. Modernized passports

do not *per se* coax a country toward the adoption of standard alphabets; however, they do ensure effcient intercommunication in the world's *scripta franca*, the Latin alphabet.

0.4.2 Machine Readable Zone

In contrast, the content of the Machine Readable Zone is highly controlled. The only characters allowed in the two lines of the MRZ are those belonging to a defined ASCII subset: (0-9, A-Z, and <). Moreover, these characters must be printed in the typeface OCR-B (OCR=Optical Character Recognition) using character and line spacings strictly defined in the 9309 standard. The adherence to these guidelines allows for unambiguous machine recognition.



Not to scale

In the Latin alphabet, most characters with diacritical marks simply have the mark dropped; some characters, however, do have special encodings to losslessly transliterate the character. The document provides a more extensive scheme for the Cyrilic and Arabic scripts, which allows nearly lossless recovery of the

original form from the MRZ content. They even provide a sample Python program for converting from the MRZ name to Unicode Arabic ([**ICAO9309**] (3.B.7.1).

Latin

The ICAO tries to account for the varying importance of diacritical marks in Latin-based scripts. Those such as the acute or grave accents, which appear over vowels mainly for the purpose of clarifying pronunciation, are simply eliminated in the MRZ. However, other characters receive recommended encoding methods These are the more "salient" diacritic characters, such as the German umlauts (ä, ö, ü) or the Spanish ñ, which in their respective languages are considered separate letters, rather than a variation on the unaccented form. The following table shows the special encodings recommended for European diacritics; all other characters simply have the mark dropped:

Unicode	National Character	Description	Recommended transliteration
00C4	Ä	A diaeresis	AE or A
00C5	Å	A ring above	AA or A
00C6	Æ	ligature AE	AE
00D1	Ñ	N tilde	N or NXX
00D6	Ö	O diaeresis	OE or O
00D8	Ø	O stroke	OE
00DC	Ü	U diaeresis	UE or UXX or U
00DE	Þ	Thorn (Iceland)	TH
00DF	ß	double S (Germany)	SS
0132	IJ	ligature IJ (Netherlands)	IJ
0152	Œ	ligature OE	OE

([**ICAO9309**] 3.6.A)

The name "Térèsa Cañón" would become CANXXON«TERESA in the MRZ. The ñ is encoded in the MRZ, while no distinction is made of the é or è. Likewise, the German name "Wilhelm Furtwängler" would become FURTWAENGLER«WILHELM (ä becomes AE). (b.4.2) Although it leaves a large set of European characters unrepresented, it would not be difficult to expand the escape sequence system to represent additional diacritical marks. (An interesting edge case would be a Spanish traveller named José Nuñenxx.)

Cyrillic

The ICAO transcription system for Cyrillic characters permits a nearly one-to-one transliteration between the MRZ and the name in the original language. The system recognizes the different values that a Cyrillic glyph might take in various languages. For example the letter Ю is transliterated as "IU", unless it is the first character of a Ukrainian name, in which case "YU" is permitted. Likewise for Щ; this is SHCH, except in Bulgarian, where it is SHT.

Arabic

For example, the Arabic name ابو بكر محمد بن زكريا الرازي would be rendered in the MRZ as ABW<BKR<MXHMD«BN<ZKRYA<ALRAZY.

While the name looks incomprehensible to a human, the encoding permits the one-to-one mapping between the MRZ and the original Arabic name. See more examples in the figure below:

B.5.9 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[7]:	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<<<<<<<<<<<<<<<<<<<<<<<<