Thesis

Gabe DeFreitas

October 25, 2018

Contents

Chapter 1

Introduction

1.1 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)

1.1.1 Names

Article 7 of the Convention on the Rights of the Child enshrines a person's inalienable right to a name: "The child shall be registered immediately after birth and shall have the right from birth to a name." [2]

Intrinsic Functions

A name is the fundamental token which identifies a person *qua* individual; thus it contains biographical information about its bearer. This can including cultural , genealogical , religious , and linguistic.

- *Genealogical*: The world's major naming practices combine a given name, which singles out particular person, with a family name or surname, which identifies that individual as belonging to some family or clan.

We can call this the *intrinsic function* of a name.

Extrinsic Functions

At the same time, a name cannot remain the sole province of its owner or family; it must facilitate interaction with the wider world as a means of address and identification. If a name affirms your status as an individual, it no less affirms your status as a citizen of your country, resident of your city, customer of your electric service provider, holder of your credit card, employee of your company, and recipient of your parking ticket. A name is worth nothing if others people in the environs cannot pronounce it, write it, or remember it. Names thus serve functions on a spectrum ranging from the expressive to the utilitarian. We can call this the *extrinsic function* of a name.

1.1.2 Computers

In 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. The ASCII standard encoding, with 127 available code points, is more than enough to represent the English language in digital form. Thus, organizations which deal only infrequently with non-English text have been slow to update their databases to Unicode standards.

Chapter 2

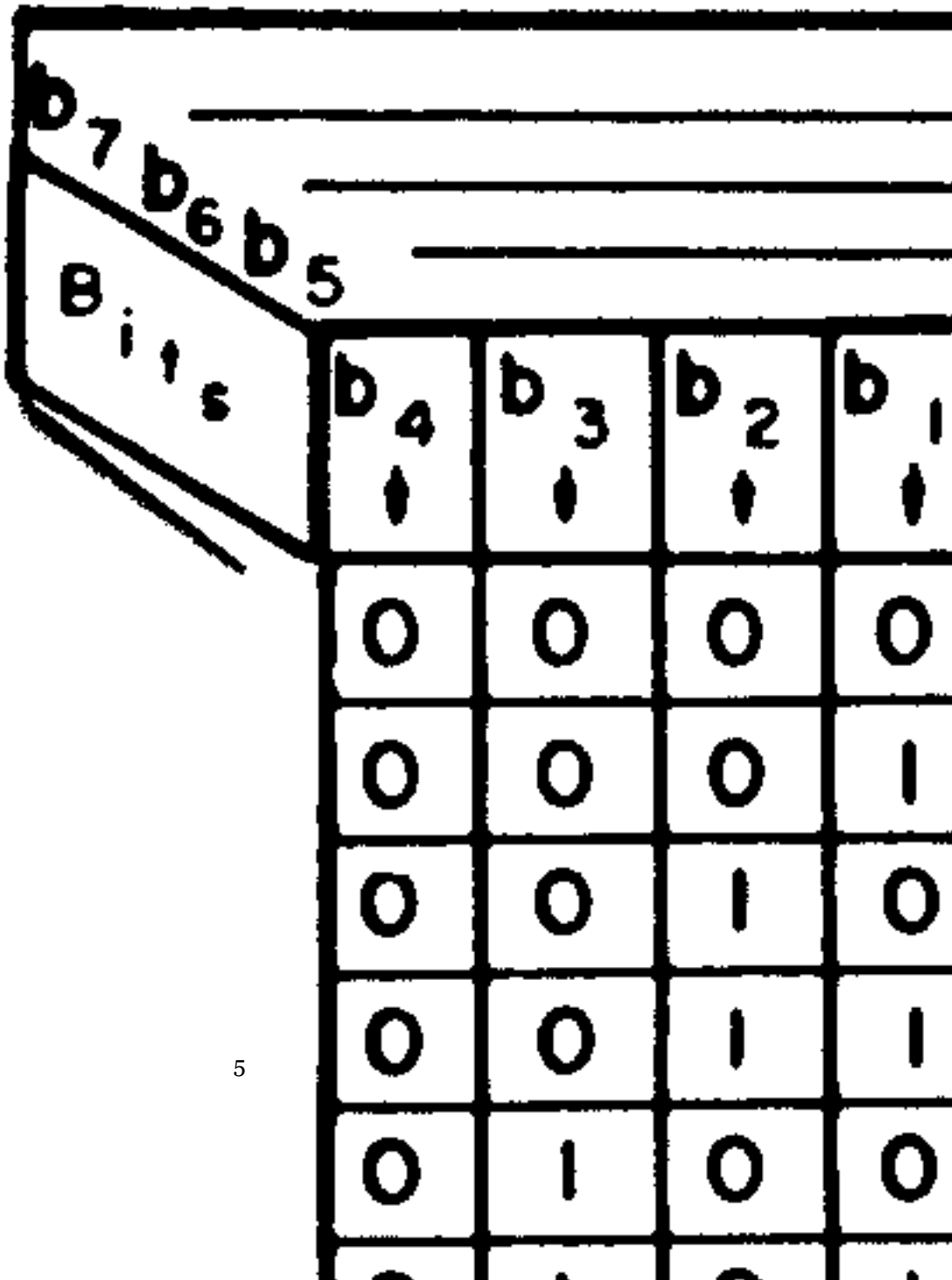
Digital Text Encoding

2.1 Unicode

2.1.1 Computers and Writing Systems

The unfortunate truth: computers can only read, write, or think in 0s and 1s. Any text that a computer deals with is ultimately stored as a number; encoding schemes describe how the computer should convert the number into a string of characters or vice versa. All computers interacting in a network must adhere to the same encoding system, lest streams of text be interpreted incorrectly and communication failures ensue.

The American Standard Code for Information Interchange was released in 1963 and quickly became the worldwide standard for digital text encoding. It was even endorsed by President Lyndon Johnson, who ordered in 1968 that all federal government computers must be ASCII-compatible. This was a natural decision for the US President, as the 128 characters in ASCII included all the necessary symbols to represent modern English text.


b7 b6 b5
Bits
b4
b3
b2
b1
0 0 0 0
0 0 0 1
0 0 1 0
0 0 1 1
0 1 0 0

The problem arose when computer technology spread beyond the labs of Silicon Valley and government offices of Washington. In other countries, there was a need to encode a vast array of linguistic forms, including diacritical marks, right-to-left scripts, and pictographic writing systems. In the 1990s, many national encoding schemes were created to represent languages other than English. Most of these were "extensions" of ASCII, meaning the original 128-character ASCII set remained intact, and non-English characters were added to the higher integer values. The only problem here is that they were not intercompatible, meaning that documents might not "play nice" when transferred between countries or languages.

Unicode and its dominant encoding scheme UTF-8 intend to alleviate these problems by collecting all of the world's glyphs into one system. The 127 characters in ASCII are represented using the same single-byte codes, which means that no conversion for existing ASCII files is necessary, and these will not become bloated in a new conversion scheme. But the encoding scheme takes advantage of modern computer's greater storage capacities to represent 1,112,064 distinct characters.

Chapter 3

Government Identification Documents

3.1 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 "" [4]. 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. Larson [4, p. 5] investigates this in his study of American naming law, finding states with such rules to include California, Massachusetts, New Hampshire, and Kansas:

3.1.1 California

Guidelines provided by the California Office of Vital Records (OVR) 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" [4].

The OVR's handbook cites Proposition 63, a 1986 ballot referendum which declared English the state's official language, with the support of 74% of voters. The initiative created Article III, Section 6 of the California Constitution, which not only explicated the status of English, but also entrusted to the state government broad powers of enforcement:

> The Legislature shall enforce this section by appropriate legislation. The Legislature and officials of the State of California shall take all steps necessary to insure that the role of English as the common language of the State of California is preserved and enhanced. The Legislature shall make no law which diminishes or ignores the role of English as the common language of the State of California.
>
> (California Constitution, Article III, Sec. 6(c))

The California Department of Public Health may see the prohibition of "non-English" characters as a natural result of Proposition 63; other government entities in California interpret the law differently. Two California state parks, Año Nuevo State Park and Montaña de Oro State Park, manage to contain the Spanish ñ in their official names, which is reflected on the parks' official webpages. [1] [5]

Likewise, the City of San José, California includes the accented é in its official name, and its Style Guide includes instructions on how to produce it digitally: "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.

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]

AB-2528 stalled in the Appropriations Committee once state agencies assigned 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.

3.1.2 Massachusetts

In Massachusetts, the "characters have to be on the standard american keyboard. So dashes and apostrophes are fine, but not accent marks and the such" [4].

3.1.3 New Hampshire

"All special characters other than an apostrophe or dash" are prohibited [4]. Technical limitations of the state's database systems prevent the inclusion of any diacritical marks.

3.1.4 Kansas

Restrictions are similar to those in Massachusetts [4].

3.2 Passports

The protocol governing machine-readable travel documents (MRTD) is Document 9309, issued by the International Civil Aviation Organization [3]. 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.

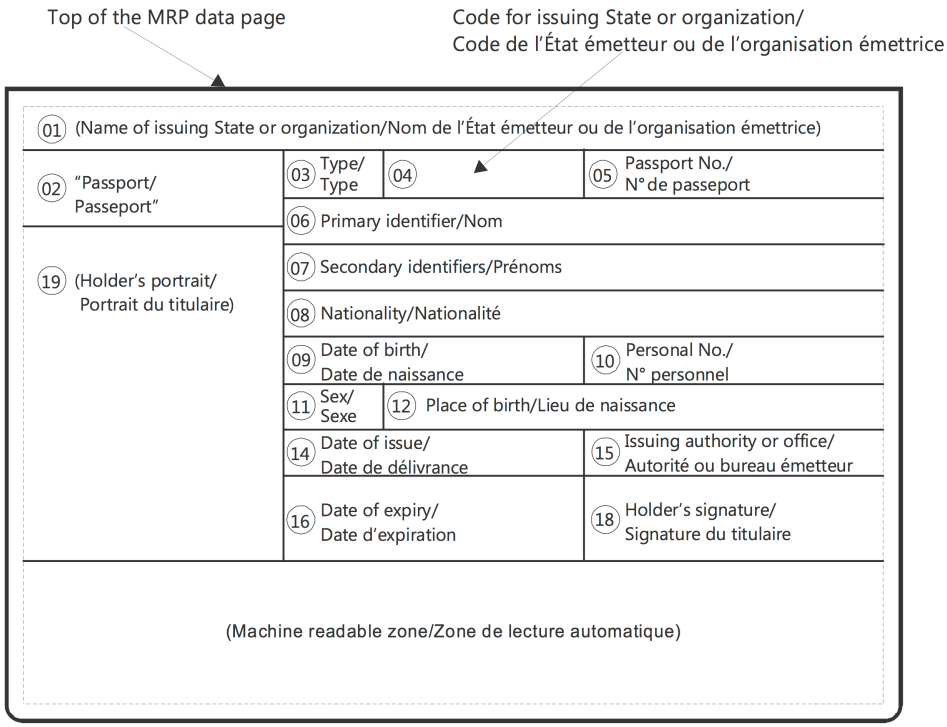
A 9309-compliant MRTD data page is divided into two sections: the Visual Inspection Zone (VIZ) and the Machine Readable Zone (MRZ).

3.2.1 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.

3.2.2 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.



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 ([3] (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	Character	Description	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

([3] 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<ÄLṘAZY.

While the name looks incomprehensible to a human, the encoding permits a 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<<<<<<<<<<<<<<<<<<<<<<

Chapter 4

Solutions

Bibliography

[1] *Año Nuevo State Park*. California Department of Parks and Recreation. URL: http://www.parks.ca.gov/?page_id=523.

[2] *Convention on the Rights of the Child*. New York: United Nations General Assembly, 1989.

[3] *Document 9309: Machine Readable Travel Documents*. International Civil Aviation Organization. Montréal, 2015.

[4] Larson, Carlton. "Naming baby: the constitutional dimensions of parental naming rights". In: *The George Washington Law Review* 80.159 (2011).

[5] *Montaña de Oro State Park*. California Department of Parks and Recreation. URL: http://www.parks.ca.gov/?page_id=592.