

IDN and the Latin Script

What We're Looking At

Repertoire

Variants

IDN

Well known scripts/languages

Chinese

中文

Arabic

العَرَبِيَّة

Cyrillic (Russian)

Росси́я

Korean

한국어

Hindi (Devanagari)

हिन्दी

Greek

ελληνικά

IDN

Less well known languages

Georgian	ქართული
Armenian	հայերեն
Cambodian (Khmer)	ភាសាខ្មែរ
Bengali (Bangla)	বাংলা
Ethiopic	አማርኛ
Gujarati	ગુજરાતી
Lao	ລາວ
Burmese (Myanmar)	မြန်မာစာ

The Latin Script is a Mess

- 210 languages
- 221 characters

Latin American and Caribbean Languages

- Spanish
- Portuguese
- Kaqchikel (Guatemala)
- Mam (Guatemala)
- Miskito (Nicaragua)
- Quechuan (Peru and Bolivia)
- Cubeo (Columbia)
- Hixkaryána (Brazil)
- Guaraní (Paraguay)
- Papiamentu (Curaçao, Bonaire and Aruba)
- Haitian Creole (Haiti)

Fonts

Times New Roman

Ariel

Courier New

a vs a

g vs g -- consider .gov vs .qov

Diacritics

Grave Accent 

Acute Accent 

Circumflex Above 

Tilde 

Macron 

Breve 

Caron 

Dot Above 

Diaeresis 

Double Acute 

Ring Above 

Hook Above

Horn 

Dot Below 

Comma Below 

Cedilla 

Ogonek 

Circumflex Below 

Macron Below 

A Small Slice

Any given language only uses a few diacritics.

- English: None!
- Spanish: Tilde, Diaeresis, Acute
- French: Cedilla, Acute, Circumflex, Grave, Diaeresis

Variants:

Is cañon noticeably different from cañon, if you don't know the Macron diacritic?

Is café noticeably different from café, if you don't know the Dot Above diacritic?

30 variations of Letter O alone!

o ȯ Ȱ ò ó ô õ ö
ø Ȫ ȫ ǒ Ǔ Ɔ ȳ Ɔ
ọ ọ̇ ọ̈ ỏ̉ ỏ̊ ỏ̋ ỏ̌ ỏ̍
ộ̎ ớ̏ ờ̐ ỡ̑ ỡ̒ ợ̓

Plus Ỗ (eth)

Variants

“What you see is what you get” . . . only works if what you see is what you *think* you see.

What you want:

- As a user, go where you expect to go
- As a registrant, having your customers/users reliably come to your site, not go somewhere else

Variants vs Confusables vs Different

Why Do You Care?

Want to register something?

- Top Level Domain names (TLDs)
 - Variants – Blocked automatically
 - Confusables – Examined by the Similarity Review Team
 - Different – Just registers
- Second Level Domain Names
(See later)

It's not a Joke

Consider this domain name:

www.test.joke

Did you notice:

That the “K” isn't just a K?

And the “J” isn't actually a J at all?

Try it bigger . . . and side-by-side and
not underlined

www.test.joke

vs

www.test.joke

Variants vs Confusables vs Different

What are they?

The “reasonably careful user”

.com

.COM

ALL CAPS

.COM

Cyrillic

.coṃ

M with Dot below

.cõm

O with Horn

.çom

C with Cedilla

.cõm

O with Circumflex and Tilde

.corn

C O R N

Cross-Script Variants

Related languages

- Cyrillic – Latin Variants 29 including:
 - Er p p P
 - Es with descender ç ç C with cedilla
- Greek – Latin Variants 18 including:
 - Nu v v V
 - Beta ß ß Sharp S
- Armenian – Latin Variants 7 including:
 - Seh g g G
 - Yiwn Լ Լ Iota

Cross-Script Variants

Generic Symbols

l	o	c	o	Latin
l	o	c		Cyrillic
l	o			Hebrew
	o			Greek
	o			Armenian
o1	o	c	o	Myanmar
		c	o	Lao
o1	o			Oriya

In-Script Variants

Schwa	ə	ə	Turned E
Iota	ɪ	ɪ	Dotless I
D with Caron	ď	đ	D with Hook
A with Breve	ă	ǎ	A with Caron
O with Diaeresis	ö	õ	O with Double Acute
Ligature AE	æ (<i>æ</i>)	œ	Ligature OE
K	k	ƀ	K with Horn

Underlining

www.example.test

www.example.test

www.example.test

www.example.test

www.example.test

www.example.test

No diacritics

L with Dot below

E with Macron below

S with Comma below

L with Circumflex below

M with Cedilla

(NOT a variant!)

Help!

During the Review Period, ***comment!***

Future Fun

Second Level Domain Names

Totally at the discretion of the Registry

- Variants? Only the ones they pick, if any
- Restrictions on variants? Only if they like

Serious economic incentives not to, of course.

Selling defensive registrations

Some Sample Variants And Confusables

- Consider the variations in a couple of letters (basic letter only):

• T: t t' t̂ t̄ ṫ ẗ t̉

• C: c c' ĉ c̄ ċ c̈ c̉

And just for variety, a vowel

• I: i i' î ī i̇ ï ỉ î ï ī

ı ĩ ı̂ ı̄ ı̇ ı̈ ı̉

So, perhaps 3 Ts, 2 Cs, and 8 Is are variants

Including confusables, we're up to: 5 Ts, 3 Cs, 14 Is

How Many Variations on a Name?

Consider just *one* domain name:

www.citi.com

- How many defensive registrations do you need, just for Variants?

www.çiti.com

C with Cedilla

www.cítí.com

I with Acute Accent

www.ciṭi.com

T with Dot Below

c i t i
2 * 8 * 3 = 48 domain names

2 * 8 * 3 * 8 = 384 domain names (if different Is are OK)

How Much Has This Happened?

- This is not a new problem

<https://www.proofpoint.com/us/resources/white-papers/domain-fraud-report>

- www.easyiet.com fraud
- But we are going to see it get a whole lot worse, thanks to the readily available list of confusable code points that we are generating

What to Do?

- Get your company involved in ICANN
 - Be an *active* stakeholder
- Get ICANN's registry contracts revised for security
 - Definitions of variants
 - Restrictions on variants
- Talk to your bank and your broker
 - They may want to get involved in self defense