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

Fonts

Times New Roman

Ariel

Courier New

a vs ɑ

g vs g -- consider .gov vs .qov

Diacritics

Grave Accent	`
Acute Accent	´
Circumflex Above	^
Tilde	~
Macron	¯
Breve	˘
Caron	ˇ
Dot Above	˙
Diaeresis	¨
Double Acute	˝
Ring Above	˚
Hook Above	
Horn	ʼ
Dot Below	.
Comma Below	,
Cedilla	¸
Ogonek	˛
Circumflex Below	^
Macron Below	ˍ

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 can̄on noticeably different from canon, if you don’t know the Macron diacritic?

Is cafė noticeably different from cafè, if you don’t know the Dot Above diacritic?

30 variations of Letter O alone!

o ǫ o̱ ò ó ô õ ö

ø ō Ơ ǒ ő ɔ ɔ̈ ɔ̱

ọ ọ̀ ọ́ ỏ ố ồ ổ ỗ

ộ ớ ờ ở ỡ ợ

Plus ð (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.įoƙe

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.ʝoƙe.test

vs

www.joke.test

Variants vs Confusables vs Different

What are they?

The “reasonably careful user”

.com	
.COM	ALL CAPS
.сом	Cyrillic
.coṃ	M with Dot below
.cơm	O with Horn
.çom	C with Cedilla
.cỗm	O with Circumflex and Tilde
.corn	C O R N

Cross-Script Variants
Related languages

- Cyrillic – Latin Variants 29 including:

Er	р	p	P
Es with descender	ҫ	ç	C with cedilla

- Greek – Latin Variants 18 including:

Nu	ν	v	V
Beta	ß	β	Sharp S

- Armenian – Latin Variants 7 including:

Seh	ց	g	G
Yiwn	ւ	ι	Iota

Cross-Script Variants
Generic Symbols

l	o	c	ɔ	Latin
ӏ	о	с		Cyrillic
ו	ס			Hebrew
	ο			Greek
	Օ			Armenian
◌ါ	ဝ	င	◌ာ	Myanmar
		ເ	ວ	Lao
◌ା	ଠ			Oriya

In-Script Variants

Schwa	ə	ǝ	Turned E
Iota	ɩ	ı	Dotless I
D with Caron	ď	ɗ	D with Hook
A with Breve	ă	ǎ	A with Caron
O with Diaeresis	ö	ő	O with Double Acute
Ligature AE	æ (*æ*)	œ	Ligature OE
K	k	ƙ	K with Horn

Underlining

www.example.test	No diacritics
www.exampḷe.test	L with Dot below
www.e̱xample.test	E with Macron below
www.example.teșt	S with Comma below
www.exampḽe.test	L with Circumflex below (NOT a variant!)
www.exam̧ple.test	M with Cedilla

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.

Some Sample Variants
And Confusables

- Consider the variations in a couple of letters (basic letter only):
- T: t ť ṱ ț ṭ ŧ ł
- C: c ć ĉ č ç ċ

And just for variety, a vowel

- I: i ĩ ı ɨ ɩ i̲ ì í î ï ī

j ǐ ɨ̃ ỉ ị

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

- 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