Comprensión de la Aceptación Universal

Mark W. Datysgeld – Governance Primer

2020

Script vs. Lenguaje

Script: conjunto de elementos base o símbolos definidos.

A, E, I, O, U ... **な** (Na) **に** (Ni), **ぬ** (Nu), **ね** (Ne), **の** (No)

Lenguaje: sistema de comunicación utilizado por una comunidad especifica.

Hola! Qué tal? ... **こんにちは、元気ですか？**

(Konnichiwa, genkidesuka?)

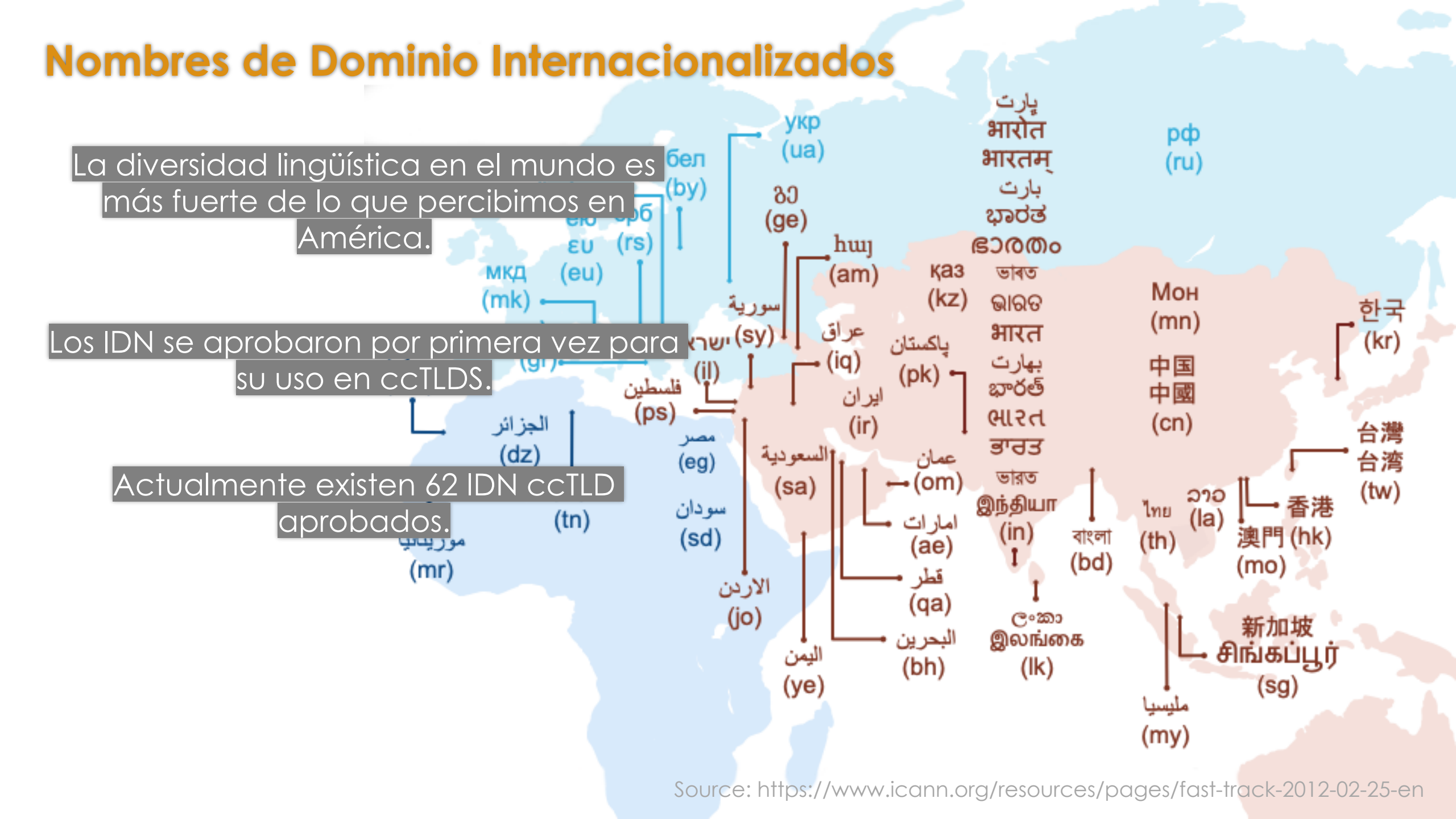
El **coreano** es un idioma que usa el script **Hangul**.

Nombres de Dominio Internacionalizados

La diversidad lingüística en el mundo es más fuerte de lo que percibimos en América.

Los IDN se aprobaron por primera vez para su uso en ccTLDS.

Actualmente existen 62 IDN ccTLD aprobados.



Source: https://www.icann.org/resources/pages/fast-track-2012-02-25-en

¿Cómo funciona?

En 2011 se hizo posible la creación de URLs completas en caracteres no ASCII. La tecnología que permite esto es **Punycode**.

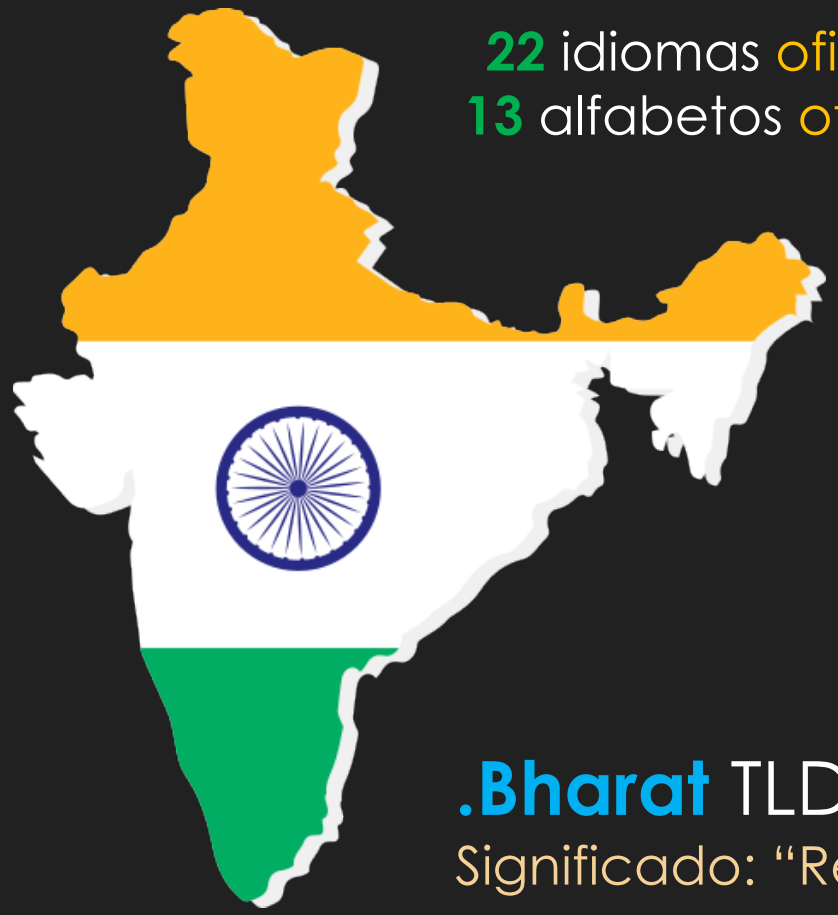
Punycode es **Unicode representado en ASCII**.

Ejemplo (U-Label): 中国移动.中国 (China Mobile)

Punycode (A-Label): xn--fiq02ib9d179b.xn--fiqs8s

Hazlo tu mismo:
https://www.punycoder.com/

IDNs: Caracteres indios



22 idiomas oficiales
13 alfabetos oficiales

IDN	Lenguaje
.भारत	Devanagari
.ভারত	Bengali
.భారత్	Telugu
.ભારત	Gujarati
.بھارت	Urdu
.இந்தியா	Tamil
.ਭਾਰਤ	Gurumukhi

.Bharat TLD
Significado: "República India"

GeoTLDs

abudhabi
africa
africa
alsace
amsterdam
aquitaine
bar
barcelona
bayern
berlin
boston
brussels
budapest
capetown
catalonia
cologne
cymru
doha
dubai
durban
frl
gent
hamburg
helsinki
ist
istanbul
joburg
koeln
kyoto
london
madrid
melbourne
miami
moscow
nagoya
nrw
nyc
okinawa
osaka
paris
quebec
rio
saarland
scot
stockholm
sydney
taipei
tata
tirol
tokyo
tui
vlaanderen
wales
wien
yokohama
zuerich
zulu
москва
佛山
广东
广州
深圳
ابوظبي

Aceptación Universal

Aceptar Validar Almacenar Procesar Presentar

- ✓ **Accept:** users can input characters from their local scripts into a text field.
- ✓ **Validate:** the software accepts the characters and recognizes them as valid.
- ✓ **Process:** when performing operations such as indexing and searching, the system can work with the characters from the user's script.
- ✓ **Store:** the software's database can receive and store the text without breaking or corrupting.
- ✓ **Display:** when fetched from the database, the information is correctly shown.

¿Por qué es difícil?

Aceptar | Validar | Almacenar | Procesar | Presentar

No se trata de cómo funciona **un solo componente**, sino de cómo **toda la estructura** de un software se ocupa de la Aceptación Universal.

Problemas: Backend

Email Address Internationalization (EAI) utiliza caracteres que no son ASCII.

- Exige que SMTP utilice la función SMTPUTF8.
- Los nombres de dominio pueden ser más largos de lo esperado.
- La escritura de derecha a izquierda (RTL) es difícil de entender y mostrar.
- Los “variant characters” se escriben de manera diferente pero significan lo mismo.
- Las pruebas deben hacerse con idiomas desconocidos para los desarrolladores.

Problemas: Código obsoleto

JavaScript Obsoleto

Muchos códigos están hechos para validar solo las letras A a la Z, números del 0 al 9 y "-" (guion), que es el estándar ASCII.

Muchos también incluyeron algunos caracteres adicionales tales como "_" (guion bajo).

```
email:function ( email ) {return /^[\w.+\-]+@ [a-zA-Z0-9-]+ (\.[a-zA-Z0-9-]+)+$/i.test(email);}
```

Otros van más allá y agregan comprobaciones redundantes para algunos TLD, así como caracteres adicionales:

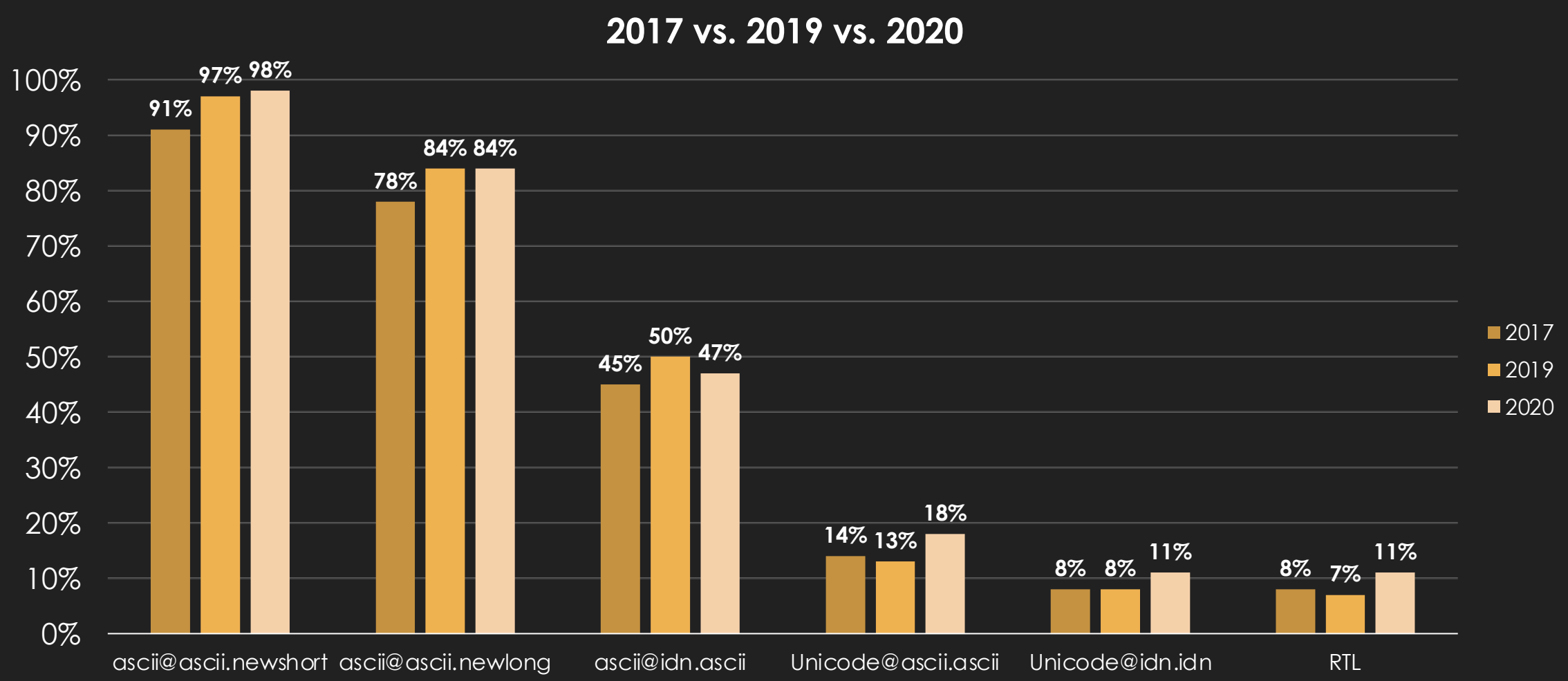
```
/^[-a-z0-9~!$%^&*_=+}{\'?]+(\.[-a-z0-9~!$%^&*_=+}{\'?]+)*@([a-z0-9_][-a-z0-9_]*(\.[-a-z09_]+)*\.(aero|arpa|biz|com|coop|edu|gov|info|int|mil|museum|name|net|org|pro|travel|mobi|[a-z][a-z])|([0-9]{1,3}\.[0-9]{1,3}\.[0-9]{1,3}\.[0-9]{1,3}))(:[0-9]{1,5})?$/i
```

Encuesta global: Premisa

Objetivo: evaluación del cumplimiento de lo UA en los 1,000 sitios Web más visitados del mundo (según Alexa) mediante el muestreo de enfoques de desarrollo para el campo del e-mail en formularios.

ascii@ascii.newshort test@test.exp	ascii@ascii.newlong test@test.example	ascii@idn.ascii test@普遍接受-测试.org
unicode@ascii.ascii 测试1@test.org	unicode@idn.idn 测试5@普遍接受-测试.世界	arabic.arabic@arabic (RTL) دون@رسيل.السعودية

Encuesta global: Resultados



Problemas: Homógrafos 1/2

¡Letras similares corresponden a diferentes códigos!

Latín	Griego	Cirilico
H	H	H
\x48	\xce\x97	\xd0\x9d

Problemas: Homógrafos 2/2

Caso 2017

Bug en Firefox y Chrome

Sitio Web de ejemplo:

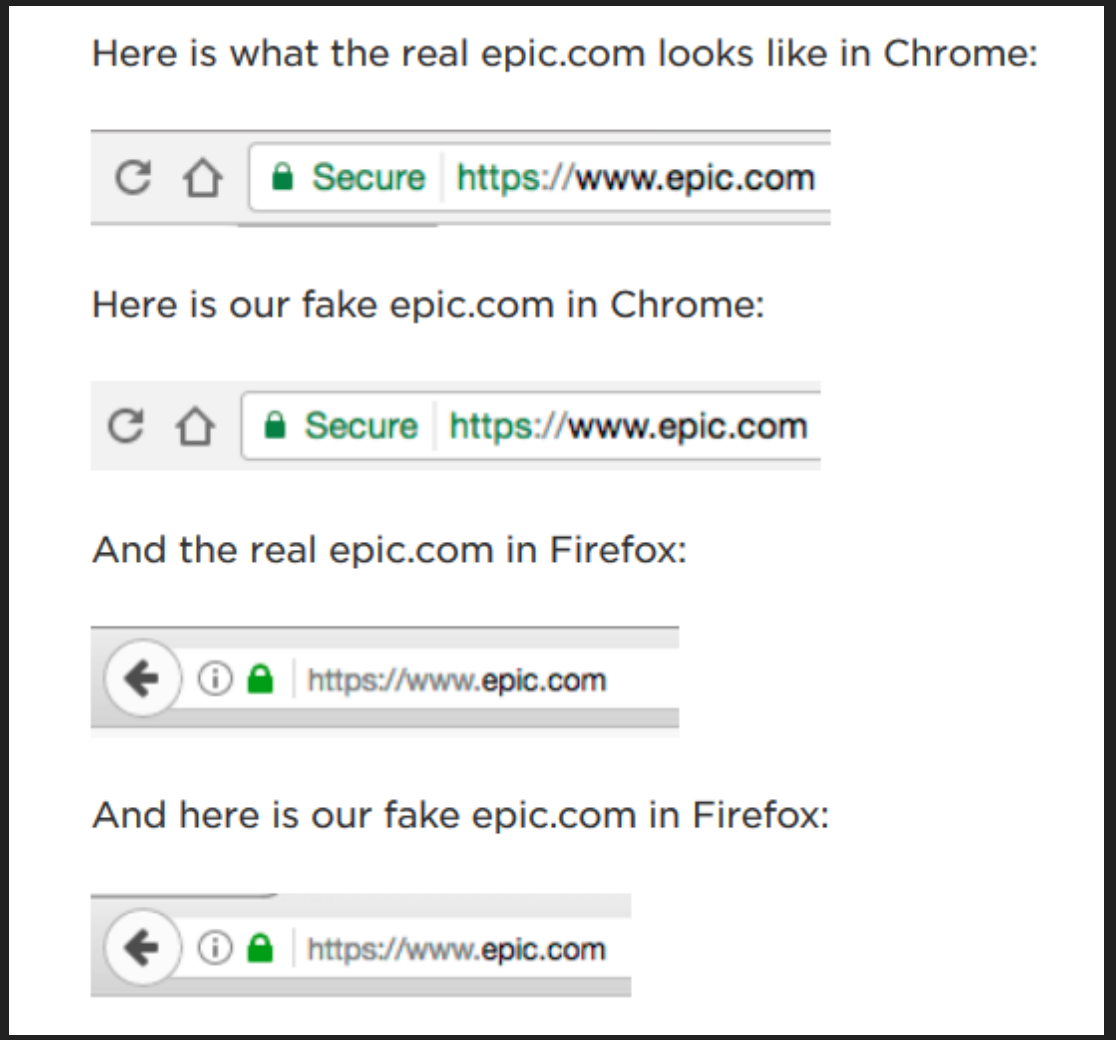
epic.com

Punycode:

https://www.xn--e1awd7f.com/

Investigación de WordFence:

https://www.wordfence.com/blog/2017/04/chrome-firefox-unicode-phishing/



¿ Estas interesado?

https://uasg.tech/

¡Gracias!

www.markwd.website