ICANN IDN guidelines & IDN Future

Marcos Sanz
sanz@denic.de

Terminology: Scripts vs Languages

ICANN Guidelines for IDN implementation

- Version 1.0 applies for registries having agreements with ICANN since June 2003
- Version 2.0 published in November 2005, big improvement compared to 1.0
- Version 2.1 published in February 2006, minor nits fixed

http://www.icann.org/topics/idn/implementation-guidelines.htm

DENIC's comments to Draft Version 2.0:
http://tinyurl.com/qqsld

Summary of the Guidelines v2.1 (I)

- Compliance to IDNA standard
- Inclusion-based approach
- Association with a domain label of:
  - one script tag, or
  - one script + one language tag, or
  - set of language tags
- All characters in a label must be from the same script
  - except if ortographies and conventions require the opposite
    - but iff no "confusable" characters in the set
- Variant tables are allowed

Summary of the Guidelines v2.1 (II)

- Allowed characters will not include:
  - Line symbol-drawing characters
  - Symbols that are neither alphanumeric nor ideographic
  - Characters with well-established functions as protocol elements
  - Punctuation marks used to indicate structure of sentences
  - Exception:
    - Punctuation marks allowed if
      - used within words,
      - essential to the language of the registration,
      - associated to a prescriptive usage context, and
      - not against any other rule.
- IDN registration in terms of both Internationalized and ACE label

Issue #1

- Problems of language-based security
  - http://www.unicode.org/reports/tr36/tr36-4.html#Language_Based_Security

Issue #2

- Script tag is useless

"□□□□" is Latin

Issue #3

- "Permissible codepoints will not include..."
  - Before: "...punctuation characters".
  - Now: "...punctuation marks *used solely to indicate the structure of sentences, but punctuation marks that are used within words may only be permitted if they [...] are essential to the language of the IDN registration, and are associated with explicit prescriptive rules about the context in which they may be used.*"

  What if no language provided (since optional)? Which are those characters? Who decides? Which rules suffice? This is even less strict than before!

Issue #4

- "Permissible codepoints will not include..."
  - Before: "...spacing characters".
  - Now: "...characters *with well-established functions as protocol elements*".

  What protocols? Which are those characters? Who decides? In the spirit of an "inclusion-based" approach, why not delivering a positive list instead of making vague exclusions and exceptions?

Issue #5

- "Visually confusable characters from different scripts will not be allowed to co-exist in a single set of permissible codepoints unless a corresponding policy and character table is clearly defined."
  - Working definition of "visually confusable" missing
  - The meaning of the sentence after "unless" is completely unclear

Issue #6

- "[All characters in a label must be from the same script] except for languages with established orthographies and conventions that require the commingled use of multiple scripts."
  - What are the exceptions and who decides?
  - What about scripts like *Common* and *Inherited*?
  - And why this restriction at all?

Mixed script confusable

- **Mixed script confusable**: Spoofing characters within more than one script and not a single script confusable.

paypal	ASCII
paypal	U+0430 Cyrillic
top	ASCII
top	U+03BF Greek

Single script confusable

- **Single script confusable**: Spoofing characters entirely within one script or using characters common across scripts (such as numbers).

Not solved

a-b	ASCII
a□b	U+0210 hyphen
dze	ASCII
□e	U+02A3 digraph
lOl	Expression of amusement
101	Binary 5

Whole script confusable

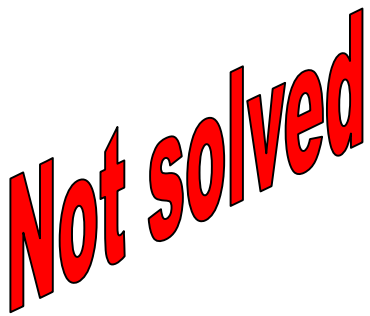
- **Whole script confusable**: Mixed script confusables where each of the strings is entirely within one script.

Not solved

caxap	Cyrillic
caxap	Latin
scope	Latin
scope	Cyrillic
SCAM	Latin
□□□□	Cherokee

"Conceptually confusable"

- **Conceptually confusable**: The labels may even be no visual confusables at all, **but the user has wrong expectations about the authenticity of the domain. THIS IS THE WEAKEST LINK.**



deutschebank.de
deutsche-bank.de
deutsche-bank-online.de
deutsche-bank-24.de
your-deutsche-bank.de
i-swear-this-is-deutsche-bank.de

Not solved either…

- Confusables for the layman: □ vs □
- Misspellings: deutcshe-bank.de
- Spoofing deeper in the DNS hierarchy
- Hunger in the world

Summary

- Improvement against previous version
- Still not clear enough:
  - open doors
  - unprecise
  - redundant
- Spoofing and scamming won't be solved with TLD registration policies. Worse: they will deliver a wrong sense of security.

sanz@denic.de