ICANN IDN Workshop
6 December 2006
São Paulo, Brazil

IDN Guidelines
Guidelines

Cary Karp
MuseDoma — dotMuseum

RFC 3490

Internationalizing Domain Names in Applications (IDNA)

March 2003

"An 'internationalized domain name' (IDN) is a domain name in which every label is an internationalized label. This implies that every ASCII domain name is an IDN (which implies that it is possible for a name to be an IDN without it containing any non-ASCII characters)."

"... every ASCII label that satisfies the [63 character] length restriction is an internationalized label."

This further implies that the term 'internationalized domain name' designates the displayed form of any domain name, regardless of what the displayed characters are.

The "stored form" of any domain name
remains restricted to ASCII,
also without regard to what the
displayed form is.

São.Paulo.tld

is stored as

xn--so-sia.paulo.tld

This displayed/stored duality
causes quite a bit of confusion
and is ***one*** of the reasons why IDN is
regarded as "so complicated".

A second reason is the difference between the implied definition of an IDN and the popular understanding of term.

A third reason is failure to appreciate the distinction between the collections of graphic symbols used for writing languages, termed

"scripts"

and the languages themselves.

The Cyrillic script is used for writing the

Bulgarian

Russian

Serbian

Ukrainian

and other languages.

The Roman script is used for writing the

Albanian

English

French

German

Portuguese

Spanish

Swahili

and many many other languages.

The ASCII character set is adequate
for the full representation of
only one of those listed here.

(And it would be naïve to assume
that it is English.)

The Universal Character Set,
also known as Unicode,
is divided into 64 different scripts
(with several more still to be added).

Not all of them are used for
writing contemporary languages,
and still fewer are ever likely to
figure in the discussion of IDN.

The present demonstration is restricted to Cyrillic and Roman scripts solely because they are convenient for the purpose.

There are other scripts that include a far greater numbers of characters and are used for multiple languages with very large speech communities.

The number of languages for which IDN support is ultimately needed can prove to be triple digit.

(The total number of languages is currently estimated at about 6,500.)

Unicode supports the automated enforcement of IDN policies that are based on script.

The extent to which it can support the useful automation of language-based policies is currently being assessed.

Regardless of the outcome,
the responsible deployment
of IDN will require registries to
adopt further policies specific to
the communities that they serve.

ICANN Guidelines for the Implementation of Internationalized Domain Names

Version 1.0

20 June 2003

Assumes that the requisite control can be based on language alone:

"In implementing the IDN standards, top-level domain registries will associate each registered internationalized domain name with one language or set of languages."

This left significant latitude for interpreting what was meant by a "set of languages".

For example, where support was provided for both German and Russian, and they were included in the same set, did that mean Cyrillic and Roman characters could appear in the same label?

If so, it would be possible to register the following ***different*** labels:

aaa (all Roman)

aaa (all Cyrillic)

aaa (CRR)

aaa (RCR)

aaa (RRC)

aaa (RCC)

aaa (CRC)

aaa (CCR)

And how does one determine
the language represented
by any of them?

Or of their numerous cousins:

áâà

aãa

äāä

ăǎă

etc.

ICANN Guidelines for the Implementation of Internationalized Domain Names

Version 2.0

7 November 2005

Assumes that the requisite control can primarily be based on script:

"In implementing the IDN standards, top-level domain registries will associate each label in a registered internationalized domain name, as it appears in their registry with a single script."

But recognizes that further modulation may be required on the basis of language:

"If greater specificity is needed, the association may be made by combining descriptors for both language and script."

And fails to eliminate the loophole:

"Alternatively, a label may be associated with a set of languages, or with more than one designator under the conditions described below."

Despite the description of

valid such conditions.

The general ban on script mixing
would shorten the list of
aaa look-alikes to only two
(pure Cyrillic and pure Roman).

But barely makes a dent
in the available combinations
of decorated letters
(eliminating only one from the list).

And for some strange reason, the Guideline that restricts the use of punctuation marks and other symbols that have no phonetic correlates has been the focus of some criticism:

Permissible code points will not include: (a) line symbol-drawing characters (as those in the Unicode Box Drawing block), (b) symbols and icons that are neither alphanumeric nor ideographic language characters, such as typographic and pictographic dingbats, (c) characters with well-established functions as protocol elements, (d) punctuation marks used solely to indicate the structure of sentences. (e) Punctuation marks that are used within words may only be permitted if they are not excluded by any of the preceding points, 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. (f) Under corresponding conditions, a single specified character may be used as a separator within a label, either by allowing the hyphen-minus to appear together with non-Latin scripts, or by designating a functionally equivalent punctuation mark from within the script.

To whatever extent it is possible to automate the reduction of the hundreds of thousands of characters in the Universal Character Set to a DNS-safer repertoire — something that can only be done on the basis of script,

the further reduction of the result
to something that is truly DNS-safe
will require the application of a
healthy amount of common sense
and responsibility —
and language considerations will weigh
heavily into the discussion about what
that level of safety can and should be.

Significant cultural sensitivity attaches both to script and to language, so the passion that may be generated during that discussion will not be appreciably reduced by excluding either concept from consideration.

The ICANN Guidelines are intended to describe TLD registry practice in a manner that is applicable in any domain name registry on any level.

And they are intended to do this in a manner that is intrinsically compelling to being implemented by those registries.

Want to help make them so?

Please post feedback on —

http://forum.icann.org/lists/idn-guidelines/