Confusing Similarity Evaluation and some of its implications
V4, July 2022

From 19 May meeting
Include reference in section 2 on lower and upper case: Included
Look at SAC060 on cases referred to by Sarmad
SAC 060 in context

1. The need for Confusing Similarity review

Avoiding confusing similarity and therefore the need to review the confusing similarity of requested IDNccTLD strings, is required to minimize the risk to the stability and security of the DNS due to user confusion by exploiting potential visual confusing similarity between domain names (eg. .PY in Latin script vs PY in Cyrillic) As such confusing similarity should therefore be minimized and mitigated. The risk of visual confusing similarity is not a technical DNS issue, but can have an adverse impact on the security and stability of the domain name system.

2. Standard & Criteria Fast Track and the 2013 proposed policy

Standard for evaluation
The proposed policy standard for evaluation to deem a selected IDNccTLD string confusing similar is the following: **If the visual appearance of the selected IDNccTLD string, both in upper and/or lower case if they are used in script of the selected IDNccTLD string, in common fonts in small sizes at typical screen resolutions is sufficiently close to one or more other strings, both in upper and/or lower case, so that it is probable that a reasonable Internet user who is unfamiliar with the script would perceive the strings to be the same or confuse one for the other.**

Notes and observations[1]: In 2013, when this standard was introduced as part of the Fast Track process, the following observations were made with respect to Common Fonts and Font Size:

> *Common Fonts*
> Fonts are a key element for determining if two strings are similar, or not, especially when considering cross script elements in the same font at the same size.
>
> Many computer applications such as web browsers use standard operating system fonts to represent URLs or search strings because these are not under the control of the of the web application, or other application, that is being accessed.
>
> As such it is possible to establish a list of common or most popular fonts which support most common scripts given these are tracked and published by various groups.
>
> Such independent ranking provides an ideal basis for selecting the fonts to be used in the comparison of strings.

[1] Notes and Observations. Additionally, and only in some instances, notes and observations from the group that developed the recommendation are included. These notes and observations are not part of the policy proposals themselves, but are provide to add depth and color to the proposals for implementation purposes and future use.

Deleted: 3
Deleted: 19 May

Font Size. Although the common number of pixels is script and font dependent, the minimum size is 9 pixels. Smaller font-sizes smaller than 9 pixels affect the readability.[2]

The Confusing similarity subgroup recommends to include these notes and observations.

3. Base for Comparison

Under the Fast Track Process and proposed IDNccTLD policy (2013) a selected IDN ccTLD string should not be deemed to be confusingly similar with:

- Any combination of two ISO 646 Basic Version (ISO 646-BV) characters[3] (letter [a-z] codes), nor
- Existing TLDs or reserved names.
- Proposed TLDs which are in process of string validation.

With the introduction of variants the **Base for Comparison** (i.e. the set of TLD strings which have to be compared on confusing similarity) will change and potentially increase significantly. This depends on category of variants that will be included in the **Base for Comparison.**

The VM group identified the following categories of variants:

I. Requested Delegatable Variants: The selected IDNccTLD string and its variants that:
   a. according to the applicable RZ-LGR are Allocatable, and
   b. are a Meaningful Representation of the name of the Territory in a Designated Language and related script, and
   c. submitted for verification as at the same time and together with the requested selected IDNccTLD string

II. Delegatable (or activatable) Variants: The selected IDNccTLD string and its variants that:
   a. according to the applicable RZ-LGR are Allocatable, and
   b. are a Meaningful Representation of the name of the Territory in a Designated Language and related script

III. Allocatable Variants

IV. Blocked Variants

In addition, and for purposes of delineating the base for comparison, note that TLDs and their variants that have been delegated or are in the process of being delegated are all or will be included in the DNS Rootzone Database and delegated in the DNS. This implies that all are different delegations (irrespective of whether they requested as a variant or selected) and should therefore always all be included the Base for Comparison. To quote from SAC120: *"From a technical perspective, two strings that are delegated in the DNS are two different delegations just like any two other domain names. Variants are no exception."*

To limit the base for comparison it is suggest to exclude blocked variants from that base. (both with respect to allocatable variants, whether or not requested or not eligible to be delegatable[4],

Scaling: Understanding the numbers

[2] http://www.w3.org/TR/CSS2/fonts.html#font-size-props, section 15.7

[3] International Organization for Standardization, "Information Technology – ISO 7-bit coded character set for information interchange," ISO Standard 646, 1991

[4] An allocatable variant may not be eligible because an eligibility criteria is NOT met, such as, but not limited to, the requirement that the variant IDNccTLD string has to be a meaningful representation of the name of a Territory.

Deleted: in size

Deleted:

Deleted: e

Deleted: for comparison

Deleted: W

Deleted: 3

Deleted: 19 May

As stated above, depending on the category of variants that will be included in the base for comparison the set of strings **against** which the requested selected IDNccTLD string and its variants need to be compared, as well the set of requested, selected IDNccTLD string and its variants, the scale of the comparison may increase significantly.

Example 1. Abu Dhabi in Arabic Script
According to ICANN's IDN Variant TLD Implementation: Appendices (see: https://www.icann.org/en/system/files/files/idn-variant-tld-appendices-25jan19-en.pdf ,page 24) the RZ- LGR for the Arabic script would degenerate 80 Variants for Abu Dhabi in Arabic script, of which 78 are blocked, 1 is valid and 1 is allocatable)
If the base for comparison would include Abu Dhabi in Arabic script:

- The base for comparison would be 1 if a selected IDNccTLD string has to be compared against Abu Dhabi in Arabic script, without variants
- The base for comparison would double if the selected IDNccTLD string would have to be compared with Abu Dhabi in Arabic script and its allocatable variant
- The base for comparison would increase 80 fold if all variants would have to compared against all variants of Abu Dhabi in Arabic script

Example 2. Pakistan in Arabic script
According to ICANN's IDN Variant TLD Implementation: Appendices, if the Arabic script RZ-LGR would be used to generate variants for "Pakistan" in the Arabic script, 1200 variants would be generated, of which 1194 are blocked and 6 are allocatable. Of the allocatable 6 variants , 3 do not represent formal or correct spellings of the name of the country in any language. Further of the 3 which represent the name of the country 1 variant is meaningful representation of the name of the country in the Designated Language, one (1) variant is poetic representation (could it be validated as name of the Pakistan?) and one (1) variant is a meaningful representation, however not in a Designated Language.

If the base for comparison would include Pakistan in Arabic script:

- The base for comparison would be 1 if a selected IDNccTLD string has to be compared against Pakistan in Arabic script, without variants
- The base for comparison would double if the selected IDNccTLD string and delegatable variants would increase two or three fold (depending on status of Poetic name for Pakistan in Urdu) if the delegatable variant would be included in the base for comparison.
- The base would increase 6 fold if all allocatable variants would be included in the comparison.
- The base for comparison would increase 1200 fold if all variants of Pakistan in the Arabic script would be included in the comparison base.

Need for delineation of the Base for Comparison
In addition to the scaling issue, the confusing similarity review may give rise to some unforeseen results and side effects if the base for comparison is not clearly demarcated. For example, if the full set of blocked variants of a requested selected IDNccTLD string should be included in the confusing similarity review as well as the blocked variants of an already delegated TLD, this could result in termination of processing of the requested and selected IDNccTLD string, because a blocked variant of the selected IDNccTLD String is confusing similar with a blocked variant of an already delegated TLD.

Formatted: Font: 11 pt
Formatted: Normal, No bullets or numbering
Deleted: the
Deleted: 3
Deleted: 19 May

It is proposed to limit the base for comparison to the following set, which has proven to be viable:

- Any combination of two ISO 646 Basic Version (ISO 646-BV) characters[5] (letter [a-z] codes), nor
- Existing TLDs or reserved names. Note this includes selected strings ang their delegated variants
- TLD strings[6] that will be requested for delegation being verified (either) in the verification process

4. Defining the Base for Comparison

It is proposed to start with the base as used under the Fast Track Process, which has proven to be viable. Accordingly, a selected IDNccTLD string should not be confusingly similar with:

- Any combination of two ISO 646 Basic Version (ISO 646-BV) characters[7] (letter [a-z] codes), nor
- Existing TLDs or reserved names
- TLD strings in the verification process (ccTLD or gTLD verification process)

With respect to requested strings, only include requested, delegatable variants in base for comparison. For purposes of the confusing similarity review, it is recommended to limit the base for comparison to **the selected IDNccTLD String and/or its requested, delegatable variants.**

The base for comparison i.e the selected IDNccTLD String and/or its requested, delegatable variants **should, be compared with:**

- **Any combination of two ISO 646 Basic Version (ISO 646-BV) characters[8] (letter [a-z] codes)**
- **Delegated TLD strings (irrespective of whether they are a variant or not) TLD strings, and their allocatable variants**
- **Verified TLD strings or labels. TLD strings or labels that have completed the applicable verification process, but have not yet been delegated, and their allocatable variants**
- **Reserved Name and**
- **Strings or labels already in one of the verification processes (IDNccTLD or IDNgTLD), and their allocatable variants**

The scale of verification is limited and pre-determined to a (small) number of requested strings that will need to be compared against the set of TLD strings that are delegated, Reserved Names or are potentially activated in the DNS, prior to or around the same time the requested, selected IDNccTLD and its requested delegatable variants may be included in the DNS and hence could give effective rise to the risks associated with confusing similarity.

[5] International Organization for Standardization, "Information Technology – ISO 7-bit coded character set for information interchange," ISO Standard 646, 1991

[6] This would include ccTLD and INDccTLD strings and their variants requested to be delegated and gTLDs and IDNgTLDs and their allocatable variants requested to be delegated.

[7] International Organization for Standardization, "Information Technology – ISO 7-bit coded character set for information interchange," ISO Standard 646, 1991

[8] International Organization for Standardization, "Information Technology – ISO 7-bit coded character set for information interchange," ISO Standard 646, 1991

Deleted: that

Deleted:

Deleted: A selected INDccTLD string should be compared with:¶

Deleted: Include

Deleted: To limit the base for comparison and to avoid unforeseen consequences the base for comparison

Formatted: Font: Not Bold

Deleted: :

Deleted: ¶
For purposes of the confusing similarity review a requested,

Formatted: Font: Not Bold

Deleted: (variant)

Deleted:

Deleted: (string

Deleted: request

Deleted:)

Deleted: or

Deleted: .

Deleted: 3

Deleted: 19 May