Confusing Similarity Evaluation and some of its implications

1. The need for Confusing Similarity review

Confusing similarity is required to minimize the risk to 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 PУ in Cyrillic), and as such should 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.

The risk of confusing similarity should be avoided because of the overarching principle of preserving the security, stability and interoperability of the DNS.

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 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.**

In 2013, when this standard was introduced, the following observations were made with respect to Common Fonts and Screen Resolutions

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.

Pixel Size. Although the common size of pixels is script and font dependent, the minimum size is 9 pixels. Smaller font-sizes smaller than 9 pixels affect the readability.[1]

It is proposed to include a note to that effect as well as clarification.

3. Base for Comparison

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

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

- Any combination of two ISO 646 Basic Version (ISO 646-BV) characters[2] (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 will change and potentially increase significantly depending on category of variants that will be included in the base for comparison.

The VM group identified the following categories of variants:

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

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

- Allocatable Variants

- Blocked Variants

Combining the Fast Track Base and the potential expansion form the inclusion of variants four sets for comparison can de delineated.

For each of these four sets a brief description is included, some scaling issues are presented and finally some questions are raised to assess whether there may some unforeseen and unwanted results.

Scaling: Understanding the numbers
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 and its allocatable variants 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

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

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 Abu Dhabi 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 is included in the comparison base.

Need for delineation of 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 the processing of the requested and selected IDNccTLD because a blocked variant of the selected IDNccTLD String is confusing similar with a blocked variant of an already delegated IDNccTLD.

Therefore, the base of comparison for the confusing similarity review needs to be re-defined and clearly delineated.

It is proposed to start with the base as used under the Fast Track Process, which has proven to be viable that a selected IDN ccTLD string should not 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
- TLD strings in the verification process

4. Sets for comparison

Requested delegatable variants included in base for comparison. Requested IDNccTLD Strings and the requested delegatable variants should be compared with:

- Any combination of two ISO 646 Basic Version (ISO 646-BV) characters[4] (letter [a-z] codes) (variant)
- TLD strings that are:

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

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

- Delegated,
- Verified (passed string request process), but not yet delegated
- Reserved Name or
- In verification process.

Scale of verification process. The scale of verification is limited and pre-determined to a (small) number of requested strings that will need to be compared with the set of TLD strings that are delegated, should be excluded (Reserved Names) or are potentially activated, prior to activation of the requested strings.

What should be the result of the review in the following cases?

- If the selected IDN ccTLD string is considered confusing similar with a delegated TLD, should further processing of selected IDNccTLDstring be terminated? Should processing of all requested delegatable variants and selected IDNccTLD strings be terminated?

- What if only (one of) the requested delegatable variant of the selected string is considered confusing similar? Should only this variant be excluded from further processing or should processing of the full set of delegatable strings be ended?

- What if a selected IDNccTLD string is considered confusing similar with a delegated variant of a TLD string, Reserved Name or TLD string in verification process, but not with a selected (IDNcc)TLD string? Should the processing of the selected IDN ccTLD string be terminated? Should processing of the full set of requested IDNccTLD be terminated?

All delegatable variants. The selected IDNccTLD string and all its Delegatable Variants (irrespective whether they are requested and un-requested) compared with

- Any combination of two ISO 646 Basic Version (ISO 646-BV) characters[5] (letter [a-z] codes) (variant)
- TLDs (and their delegatable variants, requested and unrequested in case of IDNccTLDs)that are:
  - Delegated,
  - Verified (passed string request process),
  - Reserved Name or
  - in verification process.

Scale of verification process. Scope of verification process includes all potentially delegatable variants (It may be questionable if these can be determined by a panel, as a panel is not in process to determine what is meaningful representation of the name of territory). The number of verifications is still limited, however difficult to determine.

Note that the base for comparison of this set (all delegatable IDNccTLD variant strings) can be determined by a panel, as neither the panel the panel nor ICANN is mandated to determine what is a meaningful representation of the name of the Territory in a Designated Language qnd hence what is a delegatabel variant of the selected IDNccTLD string.

What should be result?

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

- If the selected IDN ccTLD string is considered confusing similar with a delegatable variant but not requested IDNccTLD string, should the processing of all delegatable and requested IDNccTLD string be terminated?
- What if a delegatable, but not requested variant of the selected IDNccTLD string is considered to be confusing similar with a TLD, should the full set of delegatable variants and selected IDNccTLD string be excluded from further processing ? Should only the confusing similar variants of the selected IDNccTLD string be excluded?

- Should (vice versa) the delegatable but not requested variant of an already delegated INDccTLD be excluded if and when it will be requested?

- What if delegatable, but not requested variant is considered confusing similar with an already delegated TLD string?

Allocatable Variants as part of base for comparison. The selected IDNccTLD string and all its Allocatable Variants (irrespective whether they are requested and un-requested) compared with

- Any combination of two ISO 646 Basic Version (ISO 646-BV) characters[6] (letter [a-z] codes)
- TLDs (and their allocatable variants)that are:
    - Delegated
    - Verified (passed string request process),
    - Reserved Name or
    - in the verification process.

In short, all allocatable variants of the selected IDNccTLD string should be reviewed on their confusing similarity with all TLDs, Reserved Names and TLD strings/labels in a verification process and all their allocatable variants.

Scale of verification process. The verification process would include a comparison of all allocatable variants of the selected IDNccTLD string with all Delegated TLDs, Reserved Names or TLDs that are in verification process and their allocatable variants. The sets are predetermined per applied RZ-LGRs. Note that the scale of comparison will increase significantly. For example, if a selected IDNccTLD and all its allocatable variants need to be compared with Pakistan in Arabic script, the scale would increase at by a factor 6.

What should be result?

What if an allocatable, but not delegatable variant of the selected IDNccTLD string is considered confusing similar with a TLD, Reserved Name or string in verification process? Should only the confusing similar allocatable string be excluded from the process or should the selected string and all its allocatable strings be excluded from the process?

What if an allocatable, but not delegatable variant of the selected IDNccTLD string is considered confusing similar with a potential (allocatable variant) TLD, Reserved Name or string in verification process? Should only the allocatable variant of the requested IDNccTLD string be excluded? Should the full set of allocatable and selected IDNccTLD strings be excluded from processing?

---

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

Blocked variants part of base for comparison. All variants (blocked and allocatable) of selected strings are compared with all (blocked and allocatable) variants of delegated TLD, Reserved Names or in the verification process.

Scale of verification process. The scale of the verification process increases exponentially. For example, for Pakistan in the Arabic script 1200 variants are generated through the RZ-LGR. Of the total number 1194 are blocked based on the RZ-LGR. If this level of verification is preferred this would imply that all variants of the requested selected IDNccTLD string should be reviewed against all 1200 variants of Pakistan in the Arabic script.

What should be result?
What if a selected IDNccTLD string is considered confusing similar to a blocked variant of a delegated TLD, reserved name or TLD string in verification process? Should verification process be terminated?

What if a delegatable variant of a selected IDNccTLD string is considered confusing similar with a blocked variant of a delegated TLD? Should verification process be terminated?

What if allocatable IDNccTLD string is considered confusing similar to a blocked variant of a delegated TLD, reserved name or TLD string in verification process? Should verification process be terminated?

What if a blocked variant of a selected IDNccTLD string is considered confusing similar with a blocked variant of delegated TLD? Should verification process be terminated?

What if blocked variant of the selected IDNccTLD is considered confusing similar with a delegated TLD? Should processing of the full set be terminated?

Issue log, impact of scale of variants numbers in other, related areas.
Variants and the number country and territory names

1. The names (long and short version) of all Territories (see ISO3166-1, approximately 240 entries) in all Languages (depending on standard used, this can go up to approx. 7000 languages, see ISO 639-3) are excluded from gTLD process

Should variants also be taking into account? If so,
- only meaningful variants
- variants requested for delegation (subset of meaningful variants)
- Allocatable variants
- Blocked & allocatable variants