Considerations on scope of Confusing Similarity Review, including variants

Summary

Goal Confusing similarity review. The goal of the confusing similarity review is 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 РУ 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.

In SAC 060, SSAC advised ICANN (i.e the policy making bodies) that *should they decide to implement safeguards to deal with failing user expectations due to the introduction of variants, a distinction should be made between two types of failure modes: no-connection versus misconnection"*

No-connection may be a nuisance for the user, like a typo, however misconnection may result in the exploitation of the user confusion and this could be avoided though the similarity review.
With the introduction of variants one of the issues in the context of confusing similarity is to delineate the base for comparison, which is defined as the set of requested strings (Request Side) that will be compared with the set of potential visual confusingly similar strings (Comparison Side). Delineating the base for comparison is needed for reasons of :

- Scalability
- Avoiding unforeseen and/or unwanted side effects.

Under the ccNSO policy a Selected string, and its Requested Delegatable variants should not be confusingly similar with:

- Any combination of two ISO 646 Basic Version (ISO 646-BV) characters (letter [a-z] codes), nor
- Existing TLDs, which includes the already delegated variants or reserved names.
- Proposed TLDs which are in process of string validation and their requested delegatable or requested variants (however defined under the ccTLD and gTLD processes)

Introduction

At its last meeting, the CS sub-group discussed the scope of the base for comparison for the confusing similarity review. The discussion focused on the which variants, if any, to include in the comparison to assess possible confusing similarity of requested strings. Basicly the group started with assumption that the review should be based on:

- On the submission/ request side:
  - the requested label (level 1) and all allocatable variants (level 2).
- On the other side it would include:
  - Any combination of two ISO 646 Basic Version (ISO 646-BV) characters[1] (letter [a-z] codes),
  - Existing TLDs or reserved names, their allocatable (level 2) and blocked variants (level 3), and
  - Proposed TLDs which are in process of string validation, their allocatable (level 2) and blocked variants (level 3)

The scope will need to be revisited again. Starting point of this discussion is the goal of the confusing similarity review.

Goal Confusing similarity review

The agreed upon goal of the confusing similarity review is 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 PУ 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.

In SAC 060, SSAC advised ICANN (i.e the policy making bodies) that *should they decide to implement safeguards to deal with failing user expectations due to the introduction of variants, a distinction should be made between two types of failure modes: (no-connection) versus misconnection.*

- ***No-Connection (Denial of Service**): the user attempts to visit http://example.Y, reading it as being the same Uniform Resource Identifier (URI) as the http://example.X that, for example, he or she saw in an advertisement, but the connection does not work (lookup fails) because Y is either blocked, withheld, or X has no variant at all, and example.Y is not registered.*

- ***Misconnection**: the user attempts to visit http://example.Y, reading it as being the same URI as the http://example.X that, for example, he or she saw in an advertisement, but arrives at a site controlled by a registrant different to that of example.X.*

In case of no-connection, the user is frustrated and may conclude that "the Internet does not work," but no serious harm has arisen.

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

From a risk perspective: although there is a possibility (*p1*) of confusion (*C*), there is no harm (*H*)nor potential (*p2*) harm. The overall estimated impact of the risk is therefore zero *[p1*C*p2*0(=H)*].

The second case is problematic even if this effect is not the result of malicious work on the part of Y's operator or example.Y registrant. Misconnections to a perfectly legitimate site operating at example.Y present issues of possible credential compromise or other accidental disclosure of information in addition to user confusion and frustration.

From a risk perspective: there is not only a possibility (*p1*) of confusion (*C*), there is also a potential (*p2*) harm (H) to be associated with the confusing similarity. The overall estimated impact of the risk is therefore not zero *[p1*C*p2*H, whereby H>0)* and should be avoided.

Under the evolution of the Fast Track Process a joint ccNSO-SSAC working group[2] noted that in dealing with risks associated with confusing similarity *there is no general hard and fast rule with respect to the mitigation measures that should be implemented or with respect to the acceptable level of risk. It all depends very much on the circumstances, context and interplay of proposed measures and current and future risks associated with the confusing similarity of proposed strings.*

Linking these two risk categories to the goal of the confusing similarity review
Visual similarity is relevant for those situations where as a the result of visual similarity a user does not connect or misconnects. In line with the SAC060 distinction between No-connection and Misconnection a distinction should be made whereby as a result of visual confusion no-connection or a misconnection is established.

No-connection may be a nuisance for the user, like a typo, however misconnection may result in the exploitation of the user confusion and this could be avoided though the similarity review.

Scope of comparison
Taking into account the goal of the confusing similarity review, **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 PУ in Cyrillic)** the confusing similarity review is limited to avoid misconnection resulting from visual similarity of strings.

With the introduction of variants one of the issues in the context of confusing similarity is to delineate the base for comparison, which is defined as the set of requested strings (Request Side) that will be compared with the set of potential visual confusingly similar strings (Comparison Side)

[2] https://ccnso.icann.org/sites/default/files/field-attached/epsrp-final-response-17aug17-en.pdf

As a result of the introduction of variants, the potential scope of the Base for Comparison will expand exponentially. For example, as part of the confusing similarity review a selected IDNccTLD string needs to be compared with the string “Pakistan” in the Arabic script. As a result of introducing the comparison could expand to over 1200 strings (including all allocatable and blocked variants of “Pakistan” in the Arabic script). Therefore delineating the base for comparison is needed for reasons of :

- Scalability:
    - Be able to scale the review appropriately. It is expected that for the upcoming years, confusing similarity reviews have to done manually.
    - Without proper limitation, the review may become to resource intensive and/or long in duration, which may additional issues, for example around predictability.
- Avoiding unforeseen and/or unwanted side effects.
    - If the full set of blocked variants of a would be included in the Comparison Side, a requested selected IDNccTLD could be “invalid” and further processing terminated although the variant string included in the Compare Side is from another script, and co-mingling of scripts is not allowed. In other words, the comparison may include strings/labels, which are not allowed under policy.
    - If a string includes is comprised of or contains blocked variants it will never be delegated.

Comparison Side. To assess confusing similarity of strings the requested strings needs to be compared with and should not be visual similar to other strings (Comparison Side) that would include visual comparable strings from the following set:

- 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.

Delineating Scope of Request Side

The primary question to determine the scope of the Request Side Question:
Which set of variants should be taken into request side of the base for comparison?

1. Only the selected string and the requested delegatable variants?
2. The selected string and **all delegatable** variants?
3. The selected string and **all allocatable variants** of the selected string, or

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

4. The selected string and **all variants (allocatable and blocked)**.?

Proposed Request Side. The proposed policy the request side for the Base for Comparison is comprised of the:

- Selected string, and
- Requested delegatable variants (only those allocatable variants, which are a meaningful representation of the name of the territory in the designated language and related script and requested at the time of submission of the request)

Rationale

1. The IDN selection process is open and ongoing. Variants may be requested any time as long as they meet all criteria, including meaningfulness.
2. The focus should be minimizing the risk of Misconnection to minimize and/or mitigate harm.
   Abstracting from variants, if the selected string "X X" is considered confusingly similar with the string "*xx* ", which belongs to the pool of:
   - Any combination of two ISO 646 Basic Version (ISO 646-BV) characters[4] (letter [a-z] codes),
   - Existing TLDs or reserved names.
   - Proposed TLDs which are in process of string validation

   The potential misconnection results from this confusing similarity between "X X" and "*xx"* and for that reason "X X" is deemed to be invalid and processing under the policy will end.
3. From a technical point of view the selected sting "X X" and its delegatable variants should be viewed as separate TLDs. Therefore each of the requested strings should be reviewed on confusing similarity.
4. As IDNccTLD process is open and at a later stage additional variant strings may be requested (for example variants of already delegated IDNccTLD under the Fast Track process). Each of these requested variants of an already delegated selected string, should be reviewed at its own merits with respect to confusing similarity.

Delineating Scope of Comparison Side.

Re-iterating, the goal of the confusing similarity review is 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** or to paraphrase in terms of SAC 060 (*Examining the User Experience Implications of Active Variant TLDs)* the goal is to minimize the risk of Misconnection due to visual confusability of two strings.

The minimum level of the Comparison Side, before the introduction of variants, includes:

- Any combination of two ISO 646 Basic Version (ISO 646-BV) characters[5] (letter [a-z] codes), nor

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

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

- Existing TLDs or reserved names.
- Proposed TLDs which are in process of string validation.

After the introduction of the variants, the minimum set of strings in the Comparison Side, could be defined as:

- Any combination of two ISO 646 Basic Version (ISO 646-BV) characters[6] (letter [a-z] codes), nor
- Existing TLDs, which includes the already delegated variants or reserved names.
- Proposed TLDs which are in process of string validation and their requested delegatable or requested variants (however defined under the ccTLD and gTLD processes)

In other words, all strings that:

1. Should never be delegated under any existing policy (the reserved names),
2. Should always be delegatable because of other existing policy (ASCII two-letter country-code TLDs, RFC 1591)),
3. Have been delegated (existing TLDs and their delegated variants), and
4. Are in the process of validation at the time the request for the selected IDNccTLD and its requested delegatable variants was submitted. This would include the variants of the selected IDNccTLD strings and new gTLD labels and their requested variants.

Secondly, all allocatable variants could be included of all already delegated TLDs, and those which are in process.

Although, by definition allocatable variants may be requested at a later stage. The allocatable variants will need to be reviewed against all criteria, including confusing similarity and meaningfulness if they are to be delegated. By including all allocatable variants in the comparison side, the confusing similarity review could become a reservation system. Allocatable variants, which have not been requested and may never be requested could block the introduction and delegation of a selected IDNccTLD.

And again, the goal of the confusing similarity review is to minimize risk of misconnection, and therefore avoid that a requested string is potentially delegated. The goal is not to minimize or avoid Denial of Service or Non-Connection.

With respect to including the blocked variants. The arguments to exclude all allocatable variants apply even in a stronger sense.

In summary: Under the ccNSO policy a Selected string, and its Requested Delegatable variants should not be confusingly similar with:

- Any combination of two ISO 646 Basic Version (ISO 646-BV) characters (letter [a-z] codes), nor

---

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

- Existing TLDs, which includes the already delegated variants or reserved names.
- Proposed TLDs which are in process of string validation and their requested delegatable or requested variants (however defined under the ccTLD and gTLD processes)

References

https://ccnso.icann.org/sites/default/files/field-attached/epsrp-final-response-17aug17-en.pdf